CollegeBoard
connect to college success™

Research Report
No. 2007-2

Monitoring Reader Performance and DRIFT in the AP® English Literature and Composition Examination Using Benchmark Essays

Edward W. Wolfe, Carol M. Myford, George Engelhard Jr., and Jonathan R. Manalo

www.collegeboard.com

College Board Research Report No. 2007-2

Monitoring Reader Performance and DRIFT in the AP® English Literature and Composition Examination Using Benchmark Essays

Edward W. Wolfe, Carol M. Myford, George Engelhard Jr., and Jonathan R. Manalo

The College Board, New York, 2007

Edward W. Wolfe is an associate professor at Virginia Tech.

Carol M. Myford is an associate professor at the University of Illinois at Chicago.

George Engelhard Jr. is a professor at Emory University.

Jonathan R. Manalo is an assistant research scientist at ETS.

The College Board: Connecting Students to College Success

The College Board is a not-for-profit membership association whose mission is to connect students to college success and opportunity. Founded in 1900, the association is composed of more than 5,400 schools, colleges, universities, and other educational organizations. Each year, the College Board serves seven million students and their parents, 23,000 high schools, and 3,500 colleges through major programs and services in college admissions, guidance, assessment, financial aid, enrollment, and teaching and learning. Among its best-known programs are the SAT®, the PSAT/NMSQT®, and the Advanced Placement Program® (AP®). The College Board is committed to the principles of excellence and equity, and that commitment is embodied in all of its programs, services, activities, and concerns.

For further information, visit www.collegeboard.com.

Additional copies of this report (item #070482285) may be obtained from College Board Publications, Box 886, New York, NY 10101-0886, 800 323-7155. The price is $15. Please include $4 for postage and handling.

Printed in the United States of America.

Acknowledgments

We would like to acknowledge the helpful advice of Mike Linacre regarding the use of the Facets computer program to analyze the data. We are grateful to the Readers of the AP English Literature and Composition Exam and the AP Program's administrative personnel, without whose cooperation this project could never have succeeded.

The material contained herein is based on work supported by the College Board. Any opinions, findings, conclusions, or recommendations expressed are those of the authors and do not necessarily reflect the views of the College Board, the Educational Testing Service, Virginia Tech, the University of Illinois at Chicago, or Emory University.

Contents

Abstract

In this study, we investigated a variety of Reader effects that may influence the validity of ratings assigned to AP® English Literature and Composition essays. Specifically, we investigated whether Readers exhibit changes in their levels of severity and accuracy, and their use of individual scale categories over time. We refer to changes in these characteristics of Readers as Differential Reader Functioning over Time (DRIFT). Our literature review points out several weaknesses in the way Reader effects have been addressed in prior studies, and the study sought to address several of those weaknesses. The study is relevant to operational AP Readings because it addresses several existing challenges: (a) difficulties in monitoring Reader performance due to the assignment of one rating per essay; (b) difficulties in tracking changes in Reader performance over time; (c) difficulties in identifying diagnostically informative indices of DRIFT; and (d) lack of knowledge about how and when DRIFT is likely to occur during an operational AP Reading. In addition, the study suggests how to approach Reader monitoring in an automated, online reading system, should AP choose to pursue such a system in the future.

The study sought to answer research questions relating to the implications of three types of DRIFT (*differential severity*, *differential accuracy*, and *differential scale category use*) in AP English Literature and Composition essay ratings by collecting data during an operational AP Reading. Prior to the Reading, a panel of highly experienced AP Readers identified *benchmark essays* and assigned them consensus ratings. These benchmark essays were copied and distributed to AP Readers during the Reading so that the single ratings assigned to each essay could be connected via the benchmark essays. In addition, the time that each Reader began and completed rating each packet of essays during the Reading was recorded. A variety of analyses were performed for the purpose of assessing types and seriousness of various types of DRIFT, determining whether different methods of detecting DRIFT provided more or less diagnostically useful information, and determining whether connecting Readers through their ratings of the benchmark essays would result in a stronger rating design that might improve the detection of DRIFT.

Introduction

The complexity of student responses to performance assessment tasks requires expert human judgment to render measures of the quality of the responses. As a result, systematic and random error in Reader judgment (*Reader effects*) may influence the accuracy of assigned ratings. Previous research has indicated that Reader effects are common in ratings of performance assessments, that they may be diminished through Reader training and monitoring efforts, and that latent trait modeling procedures may be useful in detecting and understanding the nature of these effects (Engelhard, 1992, 1994, 1996, 2002; Engelhard and Myford, 2003; Engelhard, Myford, and Cline, 2000; Heller, Sheingold, and Myford, 1999; Linacre, Engelhard, Tatum, and Myford, 1994; Myford, Marr, and Linacre, 1996; Myford and Mislevy, 1995; Myford and Wolfe, 2003, 2004; Paulukonis, Myford, and Heller, 2000; Wolfe, 1997; Wolfe and Gitomer, 2001; Wolfe, Kao, and Ranney, 1998).

In the past, researchers studying Reader effects have tended to portray them as static characteristics of Readers. They have treated a particular Reader's effect, such as severity, as though that effect influenced every student in the same manner when, in fact, Reader effects may manifest themselves differentially as a function of other aspects of the measurement context, such as time of rating. The acronym DRIFT—Differential Reader Functioning over Time (Wolfe and Moulder, 1999, April)—refers to changes in individual Reader performance over time. In the paragraphs that follow we present examples of Reader effects that may not be static, but rather may manifest themselves at different points within a Reading. These effects may include:

- Changes in the level of severity a Reader exercises over time (i.e., becoming more or less severe as a Reading proceeds);
- Changes in a Reader's level of accuracy over time (i.e., becoming more or less accurate as a Reading proceeds); and
- Changes in a Reader's use of the scale categories over time (i.e., using the full range of categories included in the scoring guidelines at the beginning of a Reading, but later showing a restriction in the range of categories employed).

When a Reading extends over several days or weeks (as in Advanced Placement Program® [AP] Readings), certain Readers may become more severe as time progresses, exhibiting a tendency to assign lower ratings, on average, as the Reading progresses. Conversely, some Readers may become less severe, assigning higher ratings, on average, as the Reading progresses. As a consequence, the average of the ratings these particular Readers assign may vary from day to day, from morning to afternoon, or perhaps even from folder to folder. Thus, the level of severity a Reader exercises may be found to vary systematically over time (i.e., the Reader exhibits *differential severity* or *differential leniency*). A number of studies have examined whether an individual Reader's level of severity changes systematically as a Reading progresses (Bleistein and Maneckshana, 1995; Braun, 1988; Coffman and Kurfman,

1968; Hoskens and Wilson, 2001; Lumley and McNamara, 1995; Lunz and Stahl, 1990; Morgan, 1998; O'Neill and Lunz, 2000; Wilson and Case, 2000; Wood and Wilson, 1974). Results from several of these studies suggest that, in some settings, Readers do not maintain a consistent level of severity over time.

A common concern in large-scale Readings is whether Reader training procedures have adequately prepared Readers to perform the rating task. In some assessment programs, Readers participate in ongoing training throughout a Reading or are periodically recalibrated at certain points during the Reading. If these procedures are not adequate, then it is possible that the accuracy of a Reader's ratings may change as time progresses. That is, Readers may exhibit *differential accuracy* or *differential inaccuracy* over time. More specifically, some Readers may become more accurate over time as a result of gaining practice in rating student responses. On the other hand, because some scoring guidelines are complex, and because many Readings last for several days, some Readers' levels of accuracy may deteriorate over time due to fatigue. That is, Readers may become less accurate as they tire over the course of the Reading. Hence, the error in a Reader's ratings may either decrease or increase over time as a result of practice or fatigue, respectively.

DRIFT can also influence the variability of the ratings. In the case of ratings assigned during a single time period, research has shown that some Readers exhibit central tendency (i.e., overuse of the central categories of the scoring guidelines) (Engelhard, 1994; Guilford, 1954; Saal, Downey, and Lahey, 1980). As a Reading progresses, some Readers may use the central categories more frequently than they did previously. That is, they may exhibit a gradual restriction in the range of the categories they employ over time. We refer to this type of DRIFT as *differential scale category use*. The tendency for Readers to limit their ratings to the central categories after having previously made use of the full range of categories may arise due to efforts to evaluate Readers. For example, if supervisors monitor Readers and identify those who assign ratings that frequently disagree with the ratings of other Readers, then Readers may realize that they are less likely to be singled out for feedback if they avoid assigning ratings in the extreme categories and instead assign more ratings in the central categories (i.e., adopt a "play-it-safe" strategy).

Researchers studying Reader effects have largely ignored the fact that most procedures designed to detect these effects are relative measures of rating quality. That is, these measures do not enable one to evaluate the accuracy of ratings assigned, only the level of agreement attained. Thus, commonly employed Reader effect indices depict how extreme a Reader's performance is relative to the "average" Reader (Wolfe, 1998). The problem with this approach is that it focuses on Reader *agreement* and ignores the *accuracy* of the ratings. Even when all Readers use the scoring guidelines appropriately, traditional Reader effect indices will flag some Readers as exhibiting Reader effects. On the other hand, if most Readers are using the scoring guidelines inappropriately, conventional Reader effect indices will portray the best Readers as outliers without indicating the higher quality of the ratings that they assign.

Another challenge that those who study Reader effects face concerns the design of studies of Reader effects and the degree to which ratings that subsets of Readers assign are linked to one another. In some situations, those in charge of an assessment program may try to contain scoring costs by having few Readers assign ratings to each essay. In the most extreme case, only a single Reader rates each student's essay. Such a practice creates so-called *disconnected subsets* of ratings—subsets of ratings that have no student or Reader in common. The existence of connections between subsets of ratings is essential if one wishes to describe the characteristics of students or Readers without ambiguity. For example, suppose that only one Reader rates essays for a particular subset of students and that Reader does not rate any other essays for students who are outside of that subset. Further, assume that there are differences between the mean rating of that Reader for the essays for that subset of students and the mean ratings of other Readers who rated the essays for other subsets of students. In this example, it is impossible to pinpoint the cause of the differences between the mean ratings. Multiple, equally plausible hypotheses could be posited to explain the differences. For example, perhaps the various subsets were composed of students who differed in their levels of achievement (i.e., some subsets may have included more higher-achieving students than other subsets). An alternative explanation for the differences might be that the subsets of Readers may have exercised different levels of severity when rating the students' essays (i.e., there may have been more severe Readers rating certain subsets of students' essays than other subsets). In order to determine which of these two explanations to accept, one would need to implement a rating design in which all Readers would assign ratings to a common subset of students' essays so that the mean ratings for each Reader for this particular subset of students' essays could be compared to determine the degree to which the Readers differed in their levels of severity. Hence, a lack of connectivity in rating data leads to ambiguous information concerning the existence of Reader effects.

Purposes of the Study

This study seeks to extend our understanding of Reader effects and to address challenges encountered by those who are concerned about Reader effects. In the paragraphs that follow, we outline the four purposes of our study, explaining how we planned to address each of the challenges.

Challenge #1: Investigate Reader severity as a potentially dynamic and changing, rather than as a static, effect.

One purpose of this study was to determine the degree to which individual AP English Literature and Composition Readers exercised different levels of severity (or leniency) over time. If we found evidence that some Readers did indeed exhibit differential severity (or differential leniency), our plan was to identify those Readers and pinpoint at what points in the Reading they showed changes in the levels of severity they exercised.

Challenge #2: Go beyond the study of differential severity/leniency to look at other Reader effects (accuracy and scale category use) as dynamic effects.

A second purpose of the study was to determine whether there were any AP English Literature and Composition Readers who exhibited changes over time in their levels of accuracy or in their use of scale categories. If we found evidence to suggest that one or more of these dynamic effects were present for a given reader, our plan was to conduct statistical analyses to determine at what point in the Reading we could detect changes in the Reader's level of accuracy, or use of the scale categories.

Challenge #3: Investigate the accuracy of ratings assigned, not just the level of agreement attained.

A third purpose of the study was to obtain ratings from Readers in a manner that would allow us to evaluate the accuracy, rather than merely the agreement, of the ratings that AP English Literature and Composition Readers assign. In this study, we experimented with an alternative approach to monitoring Readers—an approach that may hold considerable promise as the AP Program looks to the future and the possibility of moving to an online rating system. Having Readers rate *benchmark essays* (i.e., essays that a number of highly experienced Readers have previously rated and have assigned consensus ratings) at various times each day during a Reading as a check on their accuracy is an approach that ETS has very successfully implemented in the online rating of essays for the Graduate Management Admission Test (GMAT). By reviewing the Readers' blind ratings of benchmark essays, supervisors of GMAT Readers can quickly identify those Readers whose accuracy is waning and provide needed guidance and support. In this study, we did not present Readers with essays on a computer screen; rather, we presented benchmark essays in a paper-based rating system. Introducing benchmark essays into a Reading allowed us to monitor each Reader's level of accuracy over time, comparing a Reader's performance to known standards of performance (i.e., the highly experienced Readers' consensus ratings of the benchmark essays).

Challenge #4: Investigate the comparability of various linking strategies in rating designs.

A fourth purpose of the study was to compare the detection of Reader effects using two methods for obtaining Reader connectivity in a sparse rating matrix: (a) experienced Reader-scored benchmark essays, and (b) students' performance on the multiple-choice section of the AP Examination. Monitoring Reader performance when a single Reader rates each student's essay presents a formidable challenge. Some large-scale assessment programs use a student's performance on the multiple-choice section of an exam as a criterion measure to monitor Reader accuracy. Because only a single Reader rates each student's essay in these rating designs, it is difficult to establish adequate "connections" between all Readers so that they can be fairly compared in terms of severity, internal consistency, and level of agreement.

The Reader Management System (RMS) that the AP Program employs "connects" Readers through student responses to multiple-choice items. However, for some AP Examinations, the correlation between the students' performance on the free-response and multiple-choice sections of the exam is weak. For example, in 1999, the correlation between student performance on the free-response and multiple-choice sections of the AP English Literature and Composition Exam was 0.52. The correlations of performance on the multiple-choice section with performance on each of the three separate essays were as follows: poetry analysis essay, 0.37; prose analysis, 0.44; and analytical–expository essay, 0.40 (College Board, 1999). Given these results, some may rightly question whether using performance on the multiple-choice portion is an appropriate check on the accuracy of the Readers who rated these three essays.

Another Reader monitoring practice currently employed in AP Readings is the read-behind strategy, in which Table Leaders reread selected essays to check for Reader accuracy. Because the Table Leaders must periodically review the work of the six to eight Readers at their table, the read-behind strategy is cumbersome and time consuming to implement. Perhaps more significantly, the read-behind practice does not guard against the possibility of *table drift* and may, in fact, contribute to it. That is, an entire table of Readers may alter their rating standards if the feedback they receive from their Table Leader is unique to that Table Leader and is not shared by other Table Leaders. Indeed, past research on the AP English Literature and Composition Reading suggested that "table effects" may exist (Braun, 1988).

Our study functioned as a test of the feasibility of connecting AP Readers through their ratings of common benchmark essays. We hypothesized that this approach would allow for a more direct monitoring of Reader performance and for more accurate and precise estimates

of Reader effects than the current method of connecting Readers through student responses to multiple-choice items. We say "more direct" because the current method only allows the AP Program to obtain indirect measures of each Reader's performance (i.e., an indication of whether the ratings the Reader gives students are "in sync" with the overall level of performance each student displayed on the multiple-choice section of the test). If our approach proved feasible, it would provide the AP Program with a viable option that does not rely on the multiple-choice items as the only means available to establish Reader connectivity.

Research Questions

Listed below are the eight sets of research questions that framed this investigation:

1. Are there any AP English Literature and Composition Readers whose levels of severity change as the Reading progresses (i.e., Readers who exhibit differential severity or differential leniency over time)? If there are such Readers, at what point in the Reading do these changes become apparent?
2. Do different approaches to detecting differential severity/leniency produce similar results? How comparable are the indices of differential severity/leniency obtained from different approaches? Does each approach identify the same set of Readers as exhibiting differential severity/leniency?
3. Are there any AP English Literature and Composition Readers who exhibit differential scale category use (i.e., change from using all the categories in the scoring guidelines to using fewer categories as the Reading progresses)? If so, at what point in the Reading is this Reader effect detectable? Are there some Readers who, throughout the entire Reading, do not use all the categories on the scoring guidelines?
4. Do different approaches to detecting differential scale category use in a Reader's ratings produce similar results? How comparable are indices obtained from different statistical approaches? Does each approach identify the same set of Readers as exhibiting differential scale category use?
5. Are there any AP English Literature and Composition Readers whose levels of accuracy change as the Reading progresses? If there are such Readers, at what point in the Reading do these changes become apparent?
6. Do different approaches to detecting differential accuracy/inaccuracy produce similar results? How comparable are the indices of differential accuracy/inaccuracy obtained from different approaches? Does each approach identify the same set of Readers as exhibiting differential accuracy/inaccuracy?
7. How do our depictions of DRIFT differ when we establish Reader connectivity via student responses to multiple-choice items versus via ratings of benchmark essays?
8. How do our depictions of DRIFT differ when we utilize Reader agreement indices versus when we utilize Reader accuracy indices?

Models and Indices

In a Many-Facet Rasch Measurement analysis, the natural logarithm of the odds (logit) of each transition between adjacent scale categories is represented by a parameter that depicts a facet of the measurement context (e.g., the student's level of achievement, the Reader's severity, or the scale category's challenge). In this study, the mathematical models take several forms. One of those models is the Many-Facet Rasch Model (MFRM) containing a time facet (referred to hereafter as the *Time Facet* model) (Linacre, 1989),

$$LN\left(\frac{\pi_{nrtx}}{\pi_{nrtx-1}}\right) = A_n - S_r - R_t - T_x, \quad (1)$$

where

π_{nrtx} = probability of student n being rated x by Reader r at time t,
π_{nrtx-1} = probability of student n being rated $x - 1$ by Reader r at time t,
A_n = level of achievement for student n,
S_r = severity of Reader r,
R_t = the observed performance reduction at time t, and
T_x = difficulty of the threshold between scale categories x and $x - 1$.

The Time Facet model depicts the severity of each Reader as a static characteristic across time, but it allows the mean of the ratings to vary across time. Such a model would be useful for detecting gross changes in the mean rating across all Readers as time progresses, although changes in R_t may be due to either differences in Reader severity over time, or differences in the overall achievement of students who are rated at each time.

We can also express the Time Facet model, as well as those that follow, as a probability function—a format that is useful in discussions of the analysis of model-to-data fit,

$$\pi_{nrtx} = \frac{\exp\sum_{k=0}^{x}\left(A_n - S_r - R_t - T_k\right)}{\sum_{j=0}^{m}\exp\sum_{k=0}^{j}\left(A_n - S_r - R_t - T_k\right)}, \quad (2)$$

where

j, k = counting indices with the indicated ranges, and
m = the maximum value of the scale categories beginning with the value of zero.

Facets (Linacre, 2003) is a computer program that estimates parameters for the Time Facet model (and the remaining models we describe) from the ratings that Readers assign. The program simultaneously but statistically independently estimates these parameters and then scales them onto a single linear scale. The joint calibration of Facets makes it possible to measure Reader severity on the same scale as student achievement because all facets of the rating operation are expressed in a common metric (i.e., log-odds or logits). Researchers can then make diagnostically informative comparisons among the various facets. For each element of each facet, the analysis provides a measure, a standard error (i.e., information about the precision of that measure), and fit indices (i.e., information about how well the data fit the expectations of the measurement model).

We extended the Time Facet model to determine whether individual Readers vary in their levels of severity over time by including an interaction between Readers and time in the model (referred to hereafter as the *Interaction* model),

$$LN\left(\frac{\pi_{nrtx}}{\pi_{nrtx-1}}\right) = A_n - S_r - R_t - I_{rt} - T_x, \quad (3)$$

where
I_{rt} = the deviation of the severity of Reader r at time t from that Reader's overall severity.

When implementing the Interaction model, the Facets program first estimates the parameters for the Time Facet model. Then, the computer program anchors all of the terms contained in the Time Facet model and estimates the interaction term (I_{rt}) from the residuals of the Time Facet model (Linacre, 2003). Use of the Interaction model allows us to identify individual Readers who change their levels of severity over time. We describe this process in the next section.

Similarly, we reformulated the Time Facet model to calculate a separate estimate of each Reader's level of severity at each time point (referred to hereafter as the *Separate* model),

$$LN\left(\frac{\pi_{nrtx}}{\pi_{nrtx-1}}\right) = A_n - S_{rt} - T_x, \quad (4)$$

where
S_{rt} = severity of Reader r at time t.

In the Separate model, the Facets program estimates a unique severity parameter for each Reader at each time point, making it possible to compare the estimate of a Reader's severity at a given time point to a first baseline estimate of that Reader's severity.

Indices for detecting DRIFT

Table 1 summarizes the influence of each of several Reader effects on indices that we investigated in this study. We provide definitions of these indices in the following sections, but we present the table on page 6 as an organizational framework for that discussion. Note that Table 1 contains six indices. Two of these indices, $Z_{SAI_{r_c}}$ and $t_{l_{r_c}}$, are useful for detecting changes in Reader severity or leniency over time (i.e., differential severity and differential leniency). Another index, $F_{V(X)}$, is influenced by changes in Reader severity or leniency over time and by changes in scale category use over time (i.e., differential scale category use). Yet another index, $Z_{SR-ROR_c,SR-ROR_b}$ (a standardized difference between the single Reader–rest of the Readers correlations) is influenced only by increases and decreases in Reader agreement over time (i.e., differential accuracy and differential inaccuracy, respectively). F_{fit} is influenced by both differential scale category use and differential accuracy and inaccuracy. Finally, $Z_{E-R_c,E-R_b}$ is influenced only by differential scale category use. Detailed definitions of each of these indices follow.

Methods for detecting changes in Reader severity

In this study, we examined two indices as measures of differential severity/leniency—standardized differences and interaction terms. Psychometricians routinely use the *standardized difference* as a measure of uniform differential item functioning. Raju (1988, 1990) provides a comprehensive description of its computation and interpretation. He defined the Signed Area Index (*SAI*) for the Rasch model as the difference between the item difficulty parameter estimates for two student groups. Further, he demonstrated that the *SAI* portrays the area between the item characteristic curves for a particular dichotomously scored test item, with each curve depicting the probability that a particular student group will answer the item correctly,

$$SAI_i = D_{ig} - D_{ig}, \quad (5)$$

where
SAI_i = the *SAI* for item i, and
D_{ig} = the estimated difficulty of item i for group g.

In our study, we applied the *SAI* to the comparison of the levels of severity/leniency that Readers exercised at various time points. Using the severity estimates obtained from the Separate model displayed in Equation (4), we formulated the *SAI* as

$$SAI_{rc} = S_{rc} - S_{rb}, \quad (6)$$

Table 1

Indicators of Differential Severity/Leniency, Scale Category Use, and Accuracy/Inaccuracy

	DRIFT Effect				
Index	***Differential Severity***	***Differential Leniency***	***Differential Scale Category Use***	***Increased Accuracy***	***Decreased Accuracy***
$Z_{SAI_{rc}}$	> 0.00	< 0.00	~ 0.00	~ 0.00	~ 0.00
$t_{I_{rc}}$	> 0.00	< 0.00	~ 0.00	~ 0.00	~ 0.00
$F_{V(X)}$	~ 1.00[a]	~ 1.00[a]	< 1.00[b]	~ 1.00[c]	~ 1.00[c]
$Z_{SR-ROR_c, SR-ROR_b}$	~ 0.00	~ 0.00	~ 0.00	> 0.00	< 0.00
F_{fit}	~ 1.00	~ 1.00	< 1.00[d]	< 1.00	> 1.00
$Z_{E-R_c, E-R_b}$	~ 0.00	~ 0.00	< 0.00	~ 0.00	~ 0.00

Notes: $Z_{SAI_{rc}}$ and $t_{I_{rc}}$ are used to detect changes in Reader severity or leniency over time (i.e., differential severity and differential leniency).

$F_{V(X)}$ is used to detect differential scale category use. It is influenced by changes in Reader severity or leniency over time and by changes in scale category use over time.

$Z_{SR-ROR_c, SR-ROR_b}$ is used to detect differential accuracy and differential inaccuracy. It is influenced only by increases and decreases in Reader agreement over time.

F_{fit} is influenced by both differential scale category use and differential accuracy/inaccuracy.

$Z_{E-R_c, E-R_b}$ is influenced only by differential scale category use.

Differential severity and leniency refer to decreases and increases in the Reader's average ratings over time, respectively. Differential scale category use refers to increasing use of the central categories of the scoring guidelines over time.

[a] This is true only when the Reader's severity is well matched to the distribution of student achievement (i.e., when S_{rt} or S_r is close to the value of $\overline{A}_n$).

[b] This is true under the assumption that the Readers rate randomly equivalent samples of students' essays over time.

[c] Although it is possible that differential accuracy and inaccuracy effects may cause the variance of Readers' ratings to decrease or increase due to a reduction or increase in measurement error, respectively, this effect should be minimal.

[d] This is true under the assumption that the distribution of student achievement is centrally unimodal and is well matched to the distribution of the Reader severities.

where

SAI_{r_c} = the *SAI* for Reader *r* comparing time *c* to a baseline index,

c = a comparison time index,

b = a baseline time index, and

S_{rj} = the estimated severity of Reader *r* for time *j*.

Raju (1990) also described a statistical significance test for the *SAI*, which is identical in form to the standardized difference that Wright and Masters (1982) introduced. Our formulation of that index is

$$Z_{SAI_{rc}} = \frac{SAI_{rc}}{\sqrt{SE^2_{S_{rc}} + SE^2_{S_{rb}}}}, \quad (7)$$

where

$Z_{SAI_{rc}}$ = the standardized difference index for Reader *r*, comparing the Reader's severity estimate at time *c* to the initial baseline severity estimate, and

$SE^2_{S_{rj}}$ = the variance of Reader *r*'s severity estimate for time *j*.

To test the null hypothesis that there is no difference between the Reader's baseline severity and the Reader's severity at time *c*, we compare the value of $Z_{SAI_{rc}}$ to the standard normal distribution. Others have suggested interpreting the value of the SAI itself as an effect-size indicator, with values greater than 0.50 denoting meaningfully large effects (Draba, 1977; Swaminathan and Rogers, 1990). As formulated, positive values of SAI_{r_c} and $Z_{SAI_{r_c}}$ indicate that the Reader has become more severe over time, while negative values indicate that the Reader has become less severe over time.

When we implement the Interaction model shown in Equation (3), we obtain the interaction index (I_{rc}), the other differential severity/leniency index we investigated. For this model, the parameter estimate for each Reader-by-time combination depicts the interaction or instability of the Reader's severity to that Reader's average severity. We divided the interaction index by its standard error to create a Wald *t*-test, which tests the null hypothesis that the interaction for comparison time *c* deviates from the Reader's overall severity (S_r),

$$t_{I_{rc}} = \frac{I_{rc}}{SE_{I_{rc}}}, \quad (8)$$

where

$t_{I_{rc}}$ = the Wald *t*-test comparing I_{rc} to the null value of S_r, and

$SE_{I_{rc}}$ = the standard error of I_{rc}.

When the sample size is large, one can compare the value of $t_{I_{rc}}$ to a standard normal distribution. Previous

work with the interaction bias index has indicated that it provides information that is comparable to that provided by the standardized difference (Garner and Engelhard, 1999). The interpretation of the I_{rc} and $t_{I_{rc}}$ indices is the same as for the SAI_{rc} and $Z_{SAI_{rc}}$ indices—positive values indicate increasing severity over time, while negative values indicate decreasing severity over time.

Methods for detecting changes in Reader accuracy or in the use of scale categories

Researchers using the Many-Facet Rasch Model to examine Reader effects have made few systematic efforts to determine how static, let alone dynamic, Reader effects other than severity manifest themselves. There exists some relevant literature that rationalizes the use of particular indices for detecting specific patterns in ratings (Smith, 1996; Wright, 1991, 1995) or identifies associations between patterns of ratings and particular indices (Congdon, 1998; Engelhard, 1992, 1994; Wolfe, 2004). Only a few examples exist that simulate Reader effects and examine the influence of those effects on relevant Many-Facet Rasch indices (Manalo, 2002; Myford and Wolfe, 2002, 2003; Wolfe, Chiu, and Myford, 2000; Wolfe, Moulder, and Myford, 2001). Thus, we can only speculate about how DRIFT effects, other than differential severity/leniency, may manifest themselves in Many-Facet Rasch analyses.

In this study, we examined several indices that may prove useful for identifying Readers whose use of the categories in the scoring guidelines changed over time, as well as for identifying Readers whose level of accuracy changed over time. We discuss the indices for detecting differential scale category use and for detecting differential accuracy/inaccuracy jointly in this section because some of those indices may be sensitive to both of these types of DRIFT. We begin our discussion by identifying indices that may be useful for detecting static Reader effects, and then we extend that discussion to consider how these indices may be transformed to detect DRIFT effects.

There are two indices based on the "raw ratings" Readers assign (i.e., the ratings from which MFRM parameters are estimated) that researchers can employ to detect both static and dynamic Reader effects. The first relevant index is the standard deviation (or its square, the variance) of the raw ratings. Under the somewhat questionable assumptions that Readers rate randomly equivalent samples of students' essays, and that the distributions of Reader severities and student achievement are well matched, a researcher can use the standard deviation of the ratings as an index for identifying Readers who exhibit *static central tendency* (i.e., during a given time period, the tendency to assign ratings only in the central categories of the scoring guidelines). Specifically, provided that the Reader assigns ratings with a mean near the center of the rating scale, the standard deviation of the ratings of a Reader exhibiting central tendency will be closer to zero than the standard deviation of the ratings of Readers who do not exhibit central tendency. That is, a Reader exhibiting central tendency will assign ratings that are tightly clustered around the center of the rating scale. However, we emphasize that the standard deviation of the ratings is sensitive to both central tendency and severity/leniency. That is, if a Reader assigns ratings that are in the tails of the distribution (i.e., is severe or lenient), the standard deviation of those ratings will also be small in comparison to the standard deviation of Readers who do not exhibit central tendency or severity/leniency. Hence, the standard deviation of the ratings is of somewhat limited usefulness as a sole criterion for detecting central tendency.

Another index based on raw ratings that we used in this study is the single Reader–rest of the Readers correlation (r_{SR-ROR}), which is a generalization of the Pearson correlation (Myford and Wolfe, 2003). Facets computes the generalized Pearson coefficient by pairing each rating that a particular Reader assigns with every other rating that other Readers assign to the same students. The software computes a Pearson correlation between these pairs of ratings to obtain r_{SR-ROR}.[1] The r_{SR-ROR} index summarizes the degree to which the rank ordering of students by a particular Reader is consistent with the rank ordering of those students by the rest of the Readers.[2] If a Reader ranks students in an order different from that of other Readers, the value of r_{SR-ROR} will be close to zero. Conversely, if a Reader ranks students in an order similar to that of other Readers, then the value of r_{SR-ROR} will be close to 1. Hence, the single Reader–rest of the Readers correlation may be an indicator of Reader inaccuracy, although researchers have not as yet conducted formal studies of its utility for detecting inaccurate Readers.

We also considered several diagnostic indices relevant to the various versions of the MFRM described previously. Fit indices indicate the degree to which the assigned ratings match the "expected" ratings derived from the MFRM. (An expected rating is the rating the measurement model predicts the Reader will assign the

[1] In this study, we computed r_{SR-ROR} for each Reader at each time period using the data as configured for the Separate version of the MFRM (Equation [4]).

[2] It is important to acknowledge that we have chosen to define Reader accuracy as the degree to which Readers similarly rank order students by their achievement levels. Researchers have defined this term in several ways in the past. Some have referred to our chosen definition as an *agreement* orientation, because this definition assumes that Readers assign valid ratings, on average, and that individual ratings are accurate to the degree that they jibe with the average rating that all Readers assigned (Wolfe, 1998). An alternative definition of Reader accuracy would focus on comparing ratings that individual Readers assign to an assumed-to-be-valid external measure of student achievement (Engelhard, 1996; Wolfe, 1998). We compare measures from these two frameworks in our investigation of Research Question 8.

student's essay, given the estimated level of severity the Reader exercises and the student's estimated level of achievement). Large differences between the observed and expected ratings, particularly for individual Readers, may indicate the existence of Reader effects. At the simplest level, one can examine the residual of a rating and the associated expected rating for an individual student (n) that an individual Reader (r) assigned as

$$R_{nr} = X_{nr} - E_{nr} \tag{9}$$

where

R_{nr} = the residual,

X_{nr} = the rating that Reader r assigned to student n, and

E_{nr} = the model-based expected rating of Reader r for student n,

and

$$E_{nr} = \sum_{k=0}^{m} k\pi_{nrk}, \tag{10}$$

where π_{nrk} resembles Equation (2).[3]

One can then standardize the residual to facilitate interpretation via

$$z_{R_{nr}} = \frac{R_{nr}}{\sqrt{V_{E_{nr}}}} \tag{11}$$

where

$$V_{E_{nr}} = \sum_{k=0}^{m} \left(k - E_{nr}\right)^2 \pi_{nrk} \tag{12}$$

and

$V_{E_{nr}}$ = the model variance of the observed ratings around an expected rating so that it is the statistical information associated with the expected rating.

Finally, one can average the standardized residuals across students to produce a summary statistic for each Reader that denotes the total variability of the residuals associated with that Reader (Linacre, 1989). One can perform this averaging in two ways. The unweighted mean-square fit index ($MS_{unweighted}$) is simply the average of the Reader's squared standardized residuals across all students that a Reader rated, which is mathematically equivalent to a chi-square statistic divided by its degrees of freedom,

$$MS_{unweighted} = \frac{\sum_{n=1}^{N} z_{r_{nr}}^2}{N} \tag{13}$$

where

N = the number of students.

The weighted mean-square fit index ($MS_{weighted}$), on the other hand, is an information-weighted average (across all students that a Reader rated) of the squared standardized residuals so that the ratings the Reader assigns in the highest and lowest categories of the scoring guidelines are generally weighted less heavily,

$$MS_{weighted} = \frac{\sum_{n=1}^{N} z_{R_{nr}}^2 V_{E_{nr}}}{\sum_{n=1}^{N} V_{E_{nr}}}. \tag{14}$$

Both the weighted and unweighted mean-square fit indices have an expected value of 1 and can range from 0 to ∞. For both indices, values close to 1 indicate that the observed ratings are as close to their expected ratings as would be expected, given the amount of random variability in the data. Values less than 1 indicate that the observed ratings are closer to their expected ratings than the MFRM predicts they should be when we take model-specified randomness in the data into account (i.e., *overfit*). Values greater than 1 indicate that the observed ratings are farther from their expected ratings than the model predicts (i.e., *misfit*). In the present context, Facets computes a separate weighted and unweighted mean-square fit index for each time point based on the fit of the Reader's ratings to the Separate version of the MFRM (Equation [4]).

Researchers have used the weighted and unweighted mean-square fit indices in the past as indicators of static central tendency and inaccuracy (i.e., the inability to reliably distinguish between levels of student achievement). For example, when analyzing data from a statewide writing assessment, Engelhard (1992, 1994) observed that values of these fit indices that are less than 1 tended to be associated with Reader overuse of the central categories of the rating scale. In a simulated study, Wolfe et al. (2000) determined that the two mean-square fit indices are sensitive to both central tendency and inaccuracy, and that the two indices more accurately detect inaccuracy than central tendency. However, that study also produced *inflated* fit values when the researchers simulated raters to overuse central rating categories—the opposite of what researchers have observed in operational Readings—causing others to advise caution when interpreting the values of fit indices as indicators of specific Reader effects (Myford and Wolfe, 2003).

If one considers how overuse of the central categories of scoring guidelines might influence the square of the standardized residuals, the reason for these discrepant results becomes apparent. When a Reader exhibits a central tendency effect, that Reader tends to assign more ratings in the central categories than the MFRM would predict, based on the students' levels of achievement. As a result, the squared standardized residuals for a Reader exhibiting a central tendency effect will be close to 0.00 for students who are of average achievement (assuming

[3] Note that π_{nrk} refers to the probability that Reader r will assign a rating to student n in category k. In this case, the analog of Equation (2) replaces items (subscripted with i) with Readers (subscripted with r).

that the essays are well matched to the students' levels of achievement, which is often the case), and greater than 0.00 for students who are of below (or above) average achievement. When the distribution of student achievement is unimodal, this means that most of the squared standardized residuals will be close to 0.00. In this case, one would expect both the weighted and the unweighted mean-square fit indices to be close to zero for Readers exhibiting static central tendency. On the other hand, if the distribution of student achievement is platykurtic, or is not well matched to the distribution of Reader severities, then the average of the squared standardized residuals would increase for Readers exhibiting central tendency, causing the mean-square fit indices to increase. Hence, it is clear that, in many cases, the values of the mean-square fit indices will be close to zero for Readers exhibiting static central tendency, although this will not necessarily always be the case. Regardless, mean-square fit indices should be sensitive, to some degree, to static Reader central tendency, in addition to inaccuracy.

Another index that may be useful for detecting differential scale category use and differential accuracy/inaccuracy is the correlation between the residuals and the expected ratings for a particular reader, $r_{res,exp}$. Although no simulation studies have focused on this index, Wolfe (2004, 2005) has provided a rationale for its use in detecting Reader effects. He explained that, when static central tendency occurs, a Reader's ratings for high-achieving students will be lower than the ratings that the MFRM predicts. Conversely, the Reader's ratings for low achieving students will be higher than the ratings the MFRM predicts. As a result, the scatterplot of residuals (*y*-axis) and expected ratings (*x*-axis) should show a negative slope for a Reader exhibiting central tendency. However, in the case of static inaccuracy, the value of this correlation should approach zero. Hence, the correlation between residuals and expected ratings ($r_{res,exp}$) should be useful in differentiating between a static central tendency effect and a static inaccuracy effect.

The following paragraphs describe how we transformed these indices into statistics that are sensitive to DRIFT, rather than static, effects. Recall that two raw rating indices, the standard deviation of the ratings and the single Reader–rest of the Readers correlation, may be useful as illustrative indices of DRIFT. In the case of the standard deviation of the ratings, recall that, under the assumption that Readers rate subgroups of students who are randomly equivalent with respect to their achievement levels, a Reader who moves from using all of the scale categories to using only the central categories should show a decrease in the standard deviation of his or her ratings over time. By creating a ratio of the square of these standard deviations, we can test the null hypothesis that the standard deviation of a Reader's ratings at time *t* is equal to the initial baseline standard deviation of that Reader's ratings,

$$F_{V(X)} = \frac{S^2_{X_c}}{S^2_{X_b}}, \tag{15}$$

where

$F_{V(X)}$ = an *F* test of the null hypothesis that the two standard deviations are equal,

$S^2_{X_c}$ = the variance of the ratings for a particular Reader at comparison time *c*, and

$S^2_{X_b}$ = the variance of the ratings for a particular Reader at the baseline time.

We can then compare $F_{V(X)}$ to an *F* distribution with degrees of freedom equal to the degrees of freedom for the two standard deviations. Statistically significant values of $F_{V(X)}$ that are less than 1 indicate decreases in the variability of the Reader's ratings between the comparison time *c* and the baseline time *b* that are larger than what can be accounted for by random variability—an indicator of differential scale category use (specifically, an increasing tendency to utilize the central categories of the rating scale).[4]

Similarly, we can create a statistical test to compare two single Reader–rest of the Readers correlation coefficients (r_{SR-ROR}), the other index based on ratings, to detect changes in values of that index over time. Recall that when a Reader is highly accurate in the ratings he or she assigns, the single Reader–rest of the Readers correlation should be high in comparison to a Reader who is not as accurate. Hence, a Reader who becomes more accurate over time will produce ratings that will result in a higher single Reader–rest of the Readers correlation over time. Conversely, a Reader who becomes less accurate over time will produce ratings that will result in a lower single Reader–rest of the Readers correlation over time. We can test the null hypothesis that two correlation coefficients associated with the same Reader at different times are equal to one another by standardizing the two correlations via Fisher's *Z* transformation,

$$Z_{SR-ROR} = \frac{LN\left(1 + r_{SR-ROR}\right) - LN\left(1 - r_{SR-ROR}\right).}{2} \tag{16}$$

We can then subject the standardized values to a hypothesis test to determine whether the two correlation coefficients are statistically significantly different,

$$Z_{SR-ROR_c, SR-ROR_b} = \frac{Z_{r_{SR-ROR_c}} - Z_{r_{SR-ROR_b}}}{\sqrt{\frac{1}{N_c - 3} + \frac{1}{N_b - 3}}}, \tag{17}$$

where

$Z_{r_{SR-ROR_c}}$ = the Fisher-transformed single Reader–rest of

[4] References to a criterion value of 1 in this manuscript are made under the assumption that the sample size is relatively large. The mean of an *F* distribution equals $df_{denominator}/(df_{denominator} - 2)$, so, provided the degrees of freedom associated with the *MS* that goes into the denominator of $F_{V(X)}$ or F_{fit} is greater than, say, 10, the criterion value of 1 is reasonable. The criterion value of $df_{denominator}/(df_{denominator} - 2)$ should be used if the degrees of freedom associated with the *MS* that goes into the denominator of $F_{V(X)}$ or F_{fit} is a number close to zero.

the Readers correlation at time c,

$Z_{r_{SR-ROR_b}}$ = the Fisher-transformed single Reader–rest of the Readers correlation at the baseline time b,

N_c = the number of ratings upon which $Z_{r_{SR-ROR_c}}$ is based, and

N_b = the number of ratings upon which $Z_{r_{SR-ROR_b}}$ is based.

Finally, we can compare this statistic to a standard normal distribution. Values of the index that are significantly greater than zero indicate that the single Reader–rest of the Readers correlation at comparison time c is statistically significantly higher than the correlation at the baseline time b (signaling increasing accuracy). Values of the index that are significantly less than zero indicate that the single Reader–rest of the Readers correlation at time c is statistically significantly lower than the correlation at the baseline time b (signaling decreasing accuracy).

We can also create hypothesis test statistics for the various MFRM indices described previously. For example, we can detect changes in the magnitude of the mean-square fit indices using an F statistic. Each mean-square fit index is approximately distributed as a chi-squared statistic divided by its degrees of freedom. As a result, a ratio of two mean-square fit indices should be approximately distributed as an F statistic with degrees of freedom equal to the degrees of freedom for the two mean-square statistics involved in the ratio,

$$F_{fit} = \frac{MS_c}{MS_b} \tag{18}$$

where

F_{fit} = an F test of the null hypothesis that the two mean-square fit indices are equal,

MS_c = the mean-square fit index for a particular Reader at comparison time c, and

MS_b = the mean-square fit index for a particular Reader at the baseline time b.

We can compare the value of F_{fit} to an F distribution with the degrees of freedom associated with each mean-square fit index to test the two-tailed null hypothesis of no difference between MS_c and MS_b. Statistically significant values of F_{fit} that are less than 1 indicate a decrease in mean-square fit values over time, while statistically significant values of F_{fit} that are greater than 1 indicate an increase in mean-square fit values over time.

As mentioned previously, static central tendency and inaccuracy influence mean-square fit indices. In the case of a Reader exhibiting a central tendency effect, we typically expect the value of the Reader's mean-square fit index to be less than 1, although this may not always be the case. Over time, if a Reader were to move from using the full set of categories on the scoring guidelines to overusing the central categories, then we would typically expect the value of F_{fit} to get progressively smaller. Hence, statistically significant values of F_{fit} that are less than 1 may indicate such a change in scale category use. Similarly, when a Reader assigns accurate ratings during a given time period, residuals for that Reader are close to 0.00, so the mean-square fit index is also close to 0.00. Conversely, when a Reader assigns inaccurate ratings during a given time period, residuals tend to be large, so the mean-square fit index will be greater than 1. Hence, increases in Reader accuracy over time should result in smaller mean-square fit indices over time and values of the F_{fit} ratio that are less than 1. By contrast, decreases in Reader accuracy over time should result in larger mean-square fit indices over time and values of the F_{fit} ratio that are greater than 1.

Another index that is useful for identifying cases of differential scale category use and differential accuracy is the correlation between residuals and model-based expected ratings (Wolfe, 2004). As mentioned previously, one would expect the value of the correlation between the residuals and the expected ratings to be close to zero in the case of both accuracy and inaccuracy effects. On the other hand, recall that when a Reader uses only the central categories, the residuals for that Reader are positive for low-achieving students and negative for high-achieving students. As a result, the correlation between the Reader's residuals and expected ratings will be negative. Hence, differential scale category use will result in progressively more negative Reader residuals–expected ratings correlations over time. As was true for the single Reader–rest of the Readers correlation, we can depict changes in the Reader residuals–expected ratings correlation over time via an index analogous to Equation (17),

$$Z_{E-R_c E-R_b} = \frac{Z_{r_{E-R_c}} - Z_{r_{E-R_b}}}{\sqrt{\frac{1}{N_c - 3} + \frac{1}{N_b - 3}}}. \tag{19}$$

Summary of DRIFT indices

Recall that Table 1 summarizes how one can use these indices in combination to identify various DRIFT effects. First, differential severity and leniency influence the values of $Z_{SAI_{rc}}$ and $t_{I_{rc}}$. Specifically, the values of $Z_{SAI_{rc}}$ and $t_{I_{rc}}$ will be positive when a Reader's severity increases over time, and the values will be negative when a Reader's severity decreases over time. In addition, differential severity and leniency will influence the value of $F_{V(X)}$, but not necessarily in a consistent manner across all measurement contexts. For example, assuming that the distribution of Reader severities is well matched to the distribution of student achievement (i.e., the average severity measure is approximately equal to the average achievement measure), a Reader who becomes more severe (or lenient) as time progresses will assign more ratings in the lower (or upper) categories of the scale. This causes the variance of the assigned ratings to decrease over time,

which causes the value of $F_{V(X)}$ to decrease over time. In this particular situation, differential severity or leniency would produce small values of $F_{V(X)}$. However, several other types of DRIFT (e.g., differential scale category use and differential accuracy/inaccuracy) may also influence the value of $F_{V(X)}$. As a result of this ambiguity, we focus our attention on identifying cases of differential severity and leniency using the $Z_{SAI_{rc}}$ and $t_{I_{rc}}$ indices and merely examine the relationship of $F_{V(X)}$ to the absolute values of these two primary indices.

Second, differential scale category use influences the value of $Z_{E-R_c, E-R_b}$ and may influence the values of $F_{V(X)}$ and F_{fit}. Differential scale category use seems to have the most clear-cut influence on the value of $Z_{E-R_c, E-R_b}$. Specifically, differential scale category use results in a negative correlation between expected ratings and residuals, which produces a negative value of $Z_{E-R_c, E-R_b}$. Differential scale category use may also cause the value of F_{fit} to be less than 1 under some conditions (e.g., when the distribution of student achievement is centrally unimodal and is well matched to the distribution of the Reader severities). Similarly, differential scale category use may cause the value of $F_{V(X)}$ to be less than 1, particularly when it is reasonable to assume that the samples of student essays that a given Reader rated are randomly equivalent across time. Identifying cases of differential scale category use with F_{fit} or $F_{V(X)}$ is not straightforward for several reasons. First, differential scale category use inconsistently influences these two indices. That is, depending upon the measurement context, different indices will result. Second, other DRIFT effects may influence these two indices (e.g., differential accuracy/inaccuracy may influence both indices, and differential severity/leniency may influence $F_{V(X)}$). As a result, we focus our attention on identifying cases of differential scale category use employing the $Z_{E-R_c, E-R_b}$ index and merely examine the relationship of $F_{V(X)}$ and F_{fit} with the value of $Z_{E-R_c, E-R_b}$.

Third, differential accuracy and differential inaccuracy influence the value of $Z_{SR-ROR_c, SR-ROR_b}$ and F_{fit}. Differential accuracy and differential inaccuracy may also influence the value of $F_{V(X)}$ under some measurement conditions. Differential accuracy, in the manner in which we have chosen to define that term (i.e., as agreement between Readers included in our study), will cause the correlation between ratings of different Readers to increase over time. This would cause the value of $Z_{SR-ROR_c, SR-ROR_b}$ to be positive. Conversely, differential inaccuracy will cause the correlation between ratings of different Readers to decrease over time, which causes the value of $Z_{SR-ROR_c, SR-ROR_b}$ to be negative. Differential accuracy/inaccuracy should also result in values of F_{fit} that are less than or greater than 1, respectively. Differential accuracy/inaccuracy may also influence the value of $F_{V(X)}$. However, as mentioned in the previous paragraphs, other DRIFT effects may influence the values of F_{fit} and $F_{V(X)}$, so we focus our attempts to identify cases of differential accuracy and differential inaccuracy on the $Z_{SR-ROR_c, SR-ROR_b}$ index and then examine only the relationship of this index with the values of F_{fit} and $F_{V(X)}$.

Method

Participants

Readers

Prior to the Reading, we selected 15 tables of approximately 7 Readers each to participate in the study based on location (i.e., all were located on one side of the room to make distribution of benchmark essays easier). There were 649 Readers who rated the AP English Literature and Composition essays, and 101 of these Readers participated in this study (16 percent). Table 2 summarizes the demographic characteristics for the population and for our sample of Readers. As the table shows, our sample represented the total pool quite well. In both cases, about two-thirds of the Readers were female, and a large majority was white. In addition, a little more than half of the Readers were secondary school teachers, and the average number of years of AP Reading experience was about four, although it was slightly higher in the population of Readers than in our sample.

Students

A total of 210,964 students took the 2002 AP English Literature and Composition Examination. The 101 Readers in our study rated only a subset of the essays the students wrote for Question 1 (the prose prompt). Therefore, we included in our study all the essays for Question 1 that these Readers rated (N = 51,233, about 24 percent of the essays written for this prompt). Table 3 displays the background characteristics for the population and our sample of students. As the table shows, our sample represented the total pool quite well. In both cases, about two-thirds of the students were female, and a large majority was white. In addition, the sample and the population performed comparably on the multiple-choice section of the examination.

AP® English Literature and Composition Examination

The 2002 AP English Literature and Composition Examination consisted of two sections. Section I

Table 2

Demographic Characteristics of the 2002 AP English Literature and Composition Readers

	Total Population		*Study Sample*	
	N	**%**	**N**	**%**
Gender				
Female	402	61.94	67	66.33
Male	247	38.06	34	34.66
Race/Ethnicity				
1 American Indian/ Alaskan Native	2	0.31	1	0.99
2 Black/African American	21	3.24	8	7.92
3 Mexican American/ Chicano	13	2.00	2	1.98
4 Asian/Asian American/Pacific Islander	2	0.31	2	1.98
5 Puerto Rican	0	0.00	0	0.00
6 South, Latin, Central American/other Hispanic	5	0.77	1	0.99
7 White	516	79.51	75	74.26
8 Other	12	1.85	2	1.98
Race/ethnicity not indicated	78	12.02	10	9.90
Institution				
Secondary	372	57.32	57	56.44
College	277	42.68	44	43.56
Reader Years	***Mean***	***SD***	***Mean***	***SD***
	4.13	2.89	3.71	2.54
Total	**N**		**N**	
	649		101	

Table 3

Demographic Characteristics of the 2002 AP English Literature and Composition Students

	Total Population		*Study Sample*	
	N	**%**	**N**	**%**
Gender				
Female	134,791	63.89	32,889	64.19
Male	76,173	36.11	18,344	35.81
Race/Ethnicity				
1 American Indian/ Alaskan Native	938	0.44	212	0.41
2 Black/African American	11,514	5.46	2,723	5.31
3 Mexican American/ Chicano	8,545	4.05	1,736	3.39
4 Asian/Asian American/Pacific Islander	20,930	9.92	4,355	8.50
5 Puerto Rican	1,320	0.63	307	0.60
6 South, Latin, Central American/other Hispanic	6,271	2.97	1,348	2.63
7 White	148,964	70.61	37,743	73.67
8 Other	7,050	3.34	1,609	3.14
Race/ethnicity not indicated	5,432	2.57	1,200	2.34
Weighted Multiple-Choice Score	***Mean***	***SD***	***Mean***	***SD***
	37.42	12.87	37.68	12.65
Total	**N**		**N**	
	210,964		51,233	

contained 55 multiple-choice questions. Performance on this section accounted for 45 percent of a student's total examination score. The College Board (1999) described the content and format of Section I of the examination:

> Section I requires students to read carefully four to six texts (poems or prose passages) and to answer multiple-choice questions about their content, structure, and style. Although the questions test a student's ability to construe meaning, they also require the candidate to respond to stylistic and structural features of the text (such as patterns of imagery and the use of contrast and repetition), to understand figurative language, and to identify rhetorical or poetic devices. (p. 7)

Section II consisted of three free-response questions. Performance on Section II accounted for 55 percent of a student's total examination score. The College Board (1999) described the purpose of the free-response questions:

> On the Literature exam, students are typically required to write analytical essays on both a poem and a prose text and to apply a critical generalization about literature to a specific, appropriate text of their own choosing. The essay format allows them to demonstrate skills of organization, logic, and argument to produce a personal discussion of the text. They are also free to select aspects of the passage or poem relevant to their argument and to support their point of view with pertinent evidence.
>
> Essays also allow students to demonstrate their writing skills, which include control of syntax and grammar and breadth and exactness of vocabulary as well as the elements of

composition mentioned above. Essays provide students with an opportunity to make their individual voices heard and to show the extent to which they have come to employ a mature and effective style. (p. 2)

AP English Literature and Composition Examination Process

Students took the three-hour 2002 AP English Literature and Composition Examination in May under standardized conditions in high schools and other designated testing centers. Trained AP Exam administrators proctored the examination. Students had an hour to complete Section I of the examination (the multiple-choice questions) and two hours to complete Section II (the free-response questions). Students marked their responses to the multiple-choice questions on answer sheets and wrote their essays in a test booklet. (This study focused on the scoring of the first free-response question.)

AP English Literature and Composition scoring process

Students' answer sheets were electronically scanned, and then computer scored. The computer calculated a total multiple-choice score for each student by counting the number of questions answered correctly, the number answered incorrectly, and then deducting a fraction of the incorrectly answered questions from those answered correctly to eliminate any benefit gained from random guessing.

The scoring of the students' responses to the three essay questions on the examination took place in Daytona Beach, Florida, in June 2002. Readers were nested within questions; that is, each Reader read essays for only one of the three free-response questions.

The AP English Literature and Composition Development Committee prepared a first draft of the scoring guidelines for each of the three free-response questions at the time that test developers created the questions. After students took the examination, the Committee reviewed and refined the scoring guidelines and then tested them by evaluating a randomly selected set of student essays. (See Appendix, Exhibit A1 for a copy of the scoring guidelines for the first free-response question on the examination.)

During the first morning of the Reading, the Readers participated in a training program to prepare them to score students' essays. They reviewed the scoring guidelines and read sample benchmark papers that trained Readers had previously scored. They compared and discussed the scores for the sample essays to gain an understanding of how to apply the guidelines. The Readers then received folders of photocopied essays that trained Readers had previously scored. They used the scoring guidelines to assign scores to the essays in these folders. They reviewed their scores and, when there were disagreements, engaged in discussions to compare their rationales for assigning their scores. The Readers repeated the process with additional folders of essays until each Reader demonstrated achievement in his or her use of the scoring guidelines. (See "Exam Scoring" from the *AP Technical Manual* at http://apcentral.collegeboard.com/apc/public/exam/about_exams/1994.html for a more detailed description of the processes of creating scoring guidelines and training Readers to use the scoring guidelines. See also, "Scoring the Exams" from the *Released Exam—1999 AP English Literature and Composition*, The College Board, 1999, pp. 2–4.)

During the first five days of the Reading, Readers were seated at tables to score the students' essays. When a Reader finished scoring a folder of essays, the Reader turned in the folder and received another folder of essays to evaluate. Table Leaders selectively reviewed the essay ratings that Readers seated at their table assigned (i.e., engaged in "back Reading" to monitor Reader performance). Additionally, each Table Leader received a computer printout twice a day (i.e., at lunchtime, and at the end of each day) that provided information about the ratings that each of the Readers at the table assigned during that period. Using the quality control information they obtained from back Reading and from the Reader Management System, Table Leaders identified Readers who appeared to be having difficulty employing the scoring guidelines appropriately and engaged them in retraining activities. These quality control monitoring activities continued throughout the Reading, but the Table Leaders devoted more time and effort to these activities in the first few days of the Reading to try to identify potential problems early on.

After the Readers completed the scoring, statistical analysts determined each student's AP grade. First, they calculated a multiple-choice score and a free-response score for each student. Next, they weighted the two section scores, and then used those two section scores to calculate a composite score. (Appendix, Figure A1 contains a scoring worksheet, showing the process the statistical analysts used to derive the composite scores.) Finally, they converted the composite score range to the AP grade scale, dividing the composite scores into five groups. (For details of these processes, see "Exam Scoring" from the *AP Technical Manual* at http://apcentral.collegeboard.com/apc/public/exam/about_exams/1994.html. Also, see "How the AP Grades Are Determined" from the *Released Exam—1999 AP English Literature and Composition*, The College Board, 1999, pp. 71–72.)

Collecting additional data to facilitate detection of Reader DRIFT

To answer the eight sets of research questions, we collected additional data during the AP English Literature and Composition Reading, making this Reading different from typical AP Readings. First, highly experienced AP Readers (i.e., several ETS AP staff and AP English Literature and Composition leaders including two Table Leaders, a Question Leader, and the Chief Faculty Consultant) chose and assigned consensus ratings to a set of benchmark essays prior to the operational Reading, and the 101 Readers in this study rated subsets of these benchmark essays at various points during the Reading as a check on Reader accuracy. Second, the Readers in our study recorded the time of day that they started and finished rating each folder during the operational Reading so that we could examine Reader DRIFT.

Prior to the training set selection session, we performed power analyses to determine the minimum number of benchmark essays needed to achieve reasonable sensitivity to DRIFT effects. Based on the results from our analyses, we determined that 28 benchmark essays would be needed. Several ETS AP staff and AP English Literature and Composition leaders participated in a one-day meeting in Daytona, Florida, to select 28 benchmark essays that students had written for Question 1 (i.e., the prose prompt). They chose these particular essays because they were clear examples of each point on the 9-point AP rating scale. (These benchmarks were not to be used for training Readers; rather, they were to be used in the study.) The number of benchmarks at each location on the rating scale ranged between two and four. The task for this meeting was to select and assign consensus ratings to all 28 benchmark essays. Trios of Readers first assigned ratings to each benchmark independently. Then the entire group of Readers discussed each example individually, and came to a verbal agreement concerning the consensus score.

After the highly experienced Readers selected and rated the benchmark essays, we made multiple, identical photocopies of those essays so that aides could distribute the benchmark essays to all Readers involved in the study. We divided the 28 benchmark essays into 8 overlapping sets, each containing 7 essays and having at least one essay in common with each other set. The Chief Faculty Consultant informed the Readers about the study and its purposes prior to Reader training. He explained that (a) AP staff would *not* use the data collected in the study to evaluate Reader performance, or to make decisions about which Readers would participate in the Reading the following year; (b) Table Leaders would *not* record their ratings, nor would the researchers report the ratings to the Table Leaders; (c) Readers should *not* discuss the benchmark essays with one another; (d) Readers should *not* consult the Table Leader or the Question Leader about ratings for individual benchmark essays, or ask the Table Leader or Question Leader to answer any questions they had about these essays; and (e) Readers might see the same essay more than once over the course of the Reading.

Starting the second day of the Reading, each Reader participating in the study received a folder of benchmark essays to score twice per day for four days—one folder in the morning, and one folder in the afternoon (i.e., eight "sessions"). The Study Aide and/or Table Aides brought the folders to each table during the morning and afternoon breaks, and Readers rated the benchmark essays when they returned from their break.[5] All Readers participating in the study rated the same set of benchmark essays during each session. Readers wrote their ID on the folder, and they wrote their rating for each benchmark directly on the photocopied essay. Beginning on the second day of the Reading, the Readers also recorded the date and time that they completed each folder of essays during the four-day period of the study. (Note that this applied to *only* the 101 Readers at the 15 tables included in the study, not all Readers.) Other than distributing benchmarks and documenting time indices, the Reading was no different than other AP English Literature and Composition Readings.

Analyses

Prior to describing our methods of analysis, it is important that we discuss three choices we made that influenced how we conducted our analyses. The first of these choices concerned the strategy we would use for depicting change over time. There are several strategies that an analyst could employ to depict DRIFT. One strategy would involve comparing Reader performance to an established baseline (i.e., the benchmark comparison strategy). When carrying out this strategy, the analyst would first compute indices that characterize the level of a Reader's effects during a baseline time period (e.g., Time 1). The next step would be to compute indices that characterize the level of a Reader's effects during subsequent time periods (e.g., Time 2, Time 3, etc.). Finally, the analyst would create DRIFT indices to describe the change in Reader effects between the baseline time period and each subsequent time period.

An alternative strategy for depicting DRIFT would compare a Reader's performance from a particular time period to a moving point of reference (e.g., the time point immediately preceding the one in question—for example, comparing Time 2 to Time 1, and then comparing Time

[5] In this study, Readers knew which essays were benchmark essays; it was not feasible to disguise the fact that these essays were functioning in that way, since we were using photocopies of the essays. Ideally, we would have preferred to seamlessly embed the benchmarks into the Reading so that Readers would not be aware of which essays were benchmarks. Online Readings make this feasible, but this is not possible in large-scale, paper-based Readings.

3 to Time 2). When carrying out this strategy, the analyst would calculate a series of Reader effect indices for different time periods, and then compare each index to the index calculated from the ratings the Reader assigned in the immediately preceding time period. In this way, the analyst could look for evidence of gradual change in a Reader's rating behavior over time.

We chose to use the former of these strategies when carrying out this study (i.e., the benchmark comparison strategy) because we felt that the interpretation of indices relative to a fixed (rather than moving) point was easier to communicate. However, we believe that the choice of this strategy will not limit the interpretability of our results.

A second choice relating to our analyses involved how we defined Reader accuracy. In this study, we depict Reader accuracy in terms of the degree to which the ratings of a particular Reader agree with the ratings that the other Readers in the study assigned. We describe such a framework as an *agreement* framework. The adoption of this framework implicitly assumes that, on average, Reader effects cancel each other so that the cumulative influence of Reader effects is minimal.

An alternative framework, a Reader *accuracy* framework, defines accuracy in terms of the degree to which the ratings of a particular Reader agree with a set of external measures that we assume are valid indicators of the student's level of achievement (e.g., scores on an objectively scored examination, or consensus ratings that a panel of experienced Readers assign). The adoption of this framework implicitly assumes that the external measures are indeed valid indicators of the student's level of achievement (i.e., that they are measuring the same construct as the Readers' ratings are measuring).

We chose to adopt the former (agreement) framework for our analyses, although we compare agreement and accuracy frameworks in the analyses we conducted to answer Research Questions 7 and 8. We chose this framework because of the difficulty of and expense required for establishing a reference rating against which accuracy could be measured. That is, it would be difficult to identify a reference rating that would be unanimously supported as being a "valid" rating, and obtaining that rating would likely involve an expensive rating process. We believe that the choice of this framework does limit the generalizability of our results (i.e., what may be true for the agreement framework may not necessarily also be true for the accuracy framework).

The third choice involved selecting a strategy that would enable us to connect Readers so that they could be directly compared. The AP data matrix is very sparsely connected. Generally, only one Reader rates each essay. As a result, it is difficult to compare Readers because they do not rate essays in common. We overcame this problem by having Readers rate a common set of seven essays at each time period (we call these the benchmark essays), thus creating connectivity among the ratings the Readers assigned. The creation of these connections allows us to connect ratings from each time point so that all Readers can then be calibrated onto a common underlying linear continuum. There are several ways to create these connections, and it is not clear whether one of those methods is better than the others. Hence, we compared three connection strategies for creating DRIFT indices:

1. *Experienced Reader's strategy.* This strategy "fixes" the achievement measures of students whose essays are used as benchmarks. That is, we fixed the achievement measures of the benchmark students at values derived from the consensus ratings that the highly experienced AP Table Leaders assigned. Then, we forced the achievement measures of all nonbenchmark students to conform to the location and dispersion of the distribution of the ratings of the benchmark essays. Deviations of individual Readers from this fixed location and dispersion result in bias and misfit for individual Readers.
2. *Multiple-choice strategy.* We scaled the ratings of the essays, fixing student achievement measures at values derived from their responses to the multiple-choice questions.
3. *Benchmark strategy.* We scaled the ratings of the essays, allowing the benchmark essays to serve as the links between students (i.e., we did not fix student achievement at the values that the experienced Readers' consensus ratings dictated).

We conducted our data analyses using the Facets computer program (Linacre, 2003) and SAS (SAS Institute, 1999). We describe below the strategies we employed to answer each set of research questions we posed.

1. *Are there any AP English Literature and Composition Readers whose levels of severity change as the Reading progresses (i.e., Readers who exhibit differential severity or differential leniency over time)? If there are such Readers, at what point in the Reading do these changes become apparent?*

   We estimated parameters for the *Separate* model shown in Equation and for the *Interaction* model shown in Equation (3). We then examined the $Z_{SAI_{rc}}$ and the $t_{I_{rc}}$ indices for evidence of changes in Reader severity or leniency over time.

2. *Do different approaches to detecting differential severity/leniency produce similar results? How comparable are the indices of differential severity/leniency obtained from different approaches? Does each approach identify the same set of Readers as exhibiting differential severity/leniency?*

   We compared the flag rates for DRIFT using the various indices we created to answer Question 1. We

also examined the correlations between these two sets of indices, as well as the correlations between the absolute value of these indices with $F_{V(X)}$.

3. *Are there any AP English Literature and Composition essay Readers who exhibit differential scale category use (i.e., change from using all the categories in the scoring guidelines to using fewer categories as the Reading progresses)? If so, at what point in the Reading is this Reader effect detectable? Are there some Readers who, throughout the entire Reading, do not use all the categories on the scoring guidelines?*

   We estimated parameters for the Separate model and examined the $Z_{E-R_c, E-R_b}$ index for evidence of differential scale category use. We also examined the frequency of rating scale category use over time.

4. *Do different approaches to detecting differential scale category use in a Reader's ratings produce similar results? How comparable are indices obtained from different statistical approaches? Does each approach identify the same set of Readers as exhibiting differential scale category use?*

   We compared the values of $Z_{E-R_c, E-R_b}$ to those of $F_{V(X)}$ and F_{fit}. We also compared the rates with which individual Readers were flagged using each of these indices.

5. *Are there any AP English Literature and Composition Readers whose levels of accuracy change as the Reading progresses? If there are such Readers, at what point in the Reading do these changes become apparent?*

   We estimated parameters for the Separate model and examined the $Z_{SR-ROR_c, SR-ROR_b}$ index for evidence of Reader practice and fatigue effects.

6. *Do different approaches to detecting differential accuracy/inaccuracy produce similar results? How comparable are the indices of differential accuracy/inaccuracy obtained from different approaches? Does each approach identify the same set of Readers as exhibiting differential accuracy/inaccuracy?*

   We compared the values of $Z_{SR-ROR_c, SR-ROR_b}$ to those of F_{fit} and $F_{V(X)}$. We also compared the rates with which individual Readers were flagged using each of these indices.

7. *How do our depictions of DRIFT differ when we establish Reader connectivity via student responses to multiple-choice items versus via ratings of benchmark essays?*

   We compared the values of $Z_{SAI_{rc}}$ (and F_{fit}) that we obtained when we used the benchmark and multiple-choice strategies. We sought to determine whether the use of these two different connection strategies influenced our depictions of DRIFT.

8. *How do our depictions of DRIFT differ when we utilize Reader agreement indices versus when we utilize Reader accuracy indices?*

   We compared the values of $Z_{SAI_{rc}}$ (and F_{fit}) that we obtained when we used the multiple-choice and benchmark strategies to the values of $Z_{SAI_{rc}}$ (and F_{fit}) that we obtained when we used the experienced Reader's strategy. Our goal was to determine which one of these three connection strategies provided the best depiction of the accuracy of Readers' ratings in a DRIFT context.

Results

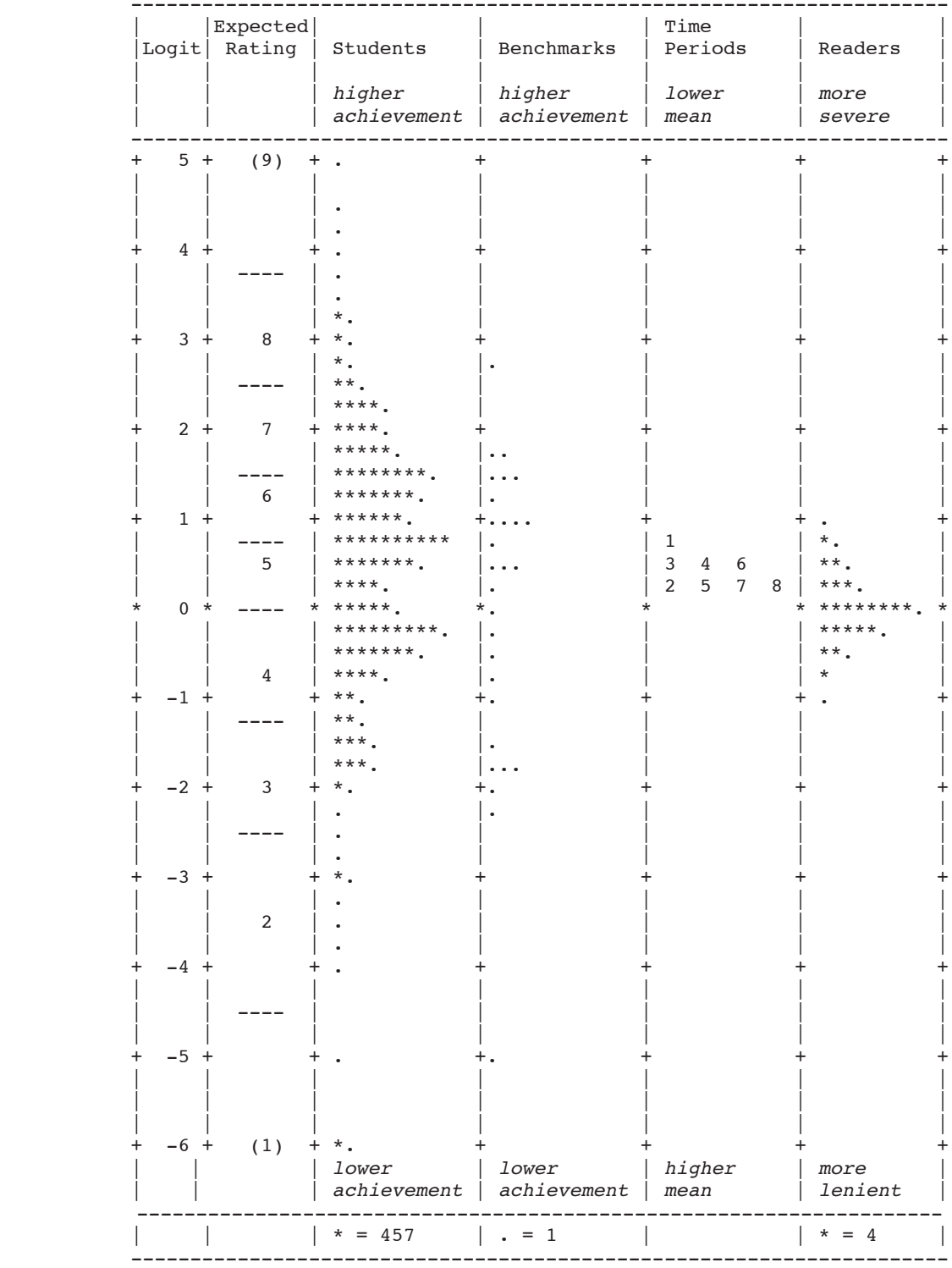
We begin our discussion of the results by presenting the indices we obtained when we analyzed the rating data using the Time Facet model (Equation [2]) with these data. Figure 1 displays the map of all measurement facets represented in this model.

The leftmost column shows the values of the equal interval logit scale. The Facets computer program calibrates the Readers, students, benchmarks, time periods, and the rating scale so that all facets are positioned on this scale, creating a single frame of reference for interpreting the results of the analysis. The remaining columns display the location of each element (e.g., each student, reader, benchmark, time period) of each facet in the Time Facet model.

The second column of Figure 1 ("Expected Rating") displays the rating scale structure mapped onto the logit scale. As shown in the second column of Table 4, the distribution of ratings was unimodal (mode = 4) for the prose essay, and Readers assigned fewer ratings in the extreme categories than in the middle categories of the 9-point scale. The third column of Table 4 provides evidence that the Readers applied the rating scale categories in the way that those who developed the scale intended the categories to be used. Specifically, the average measure of students assigned to each category increases with the categories themselves.[6] This is also true of the rating scale category thresholds, which are shown in the fourth column of Table 4.[7] The categories are fairly widely dispersed, and they are fairly precisely measured (as evidenced by the small

[6] If the scale is functioning as intended, then the average measures should increase in magnitude as the rating scale categories increase. When this pattern is borne out in the data, the results suggest that students with higher ratings are indeed exhibiting more of the construct being measured (i.e., English achievement) than are students with lower ratings.

[7] If the scale category thresholds do not increase in value, then these disordered thresholds can muddle the interpretation of the categories, since one or more of the categories are never more probable to be assigned.



Note: "Expected Rating" refers only to student and benchmark achievement. The scales of the time and Reader calibrations are reversed (i.e., for time period, the lower the logit measure, the higher the mean; for Readers, the lower the logit measure, the more lenient the Reader).

Figure 1. Time Facet map.

Table 4

Rating Scale Category Statistics for the Time Facet Model

Category	*Percent*	*Average Measure*	*Category Threshold*	*SE*	$MS_{unweighted}$
1	4%	–4.43			0.9
2	10%	–2.46	–4.17	0.09	0.9
3	15%	–1.55	–2.40	0.05	1.0
4	24%	–.41	–1.45	0.04	1.0
5	20%	0.37	0.19	0.04	1.1
6	14%	0.79	0.93	0.04	1.1
7	8%	1.17	1.61	0.05	1.2
8	4%	1.64	2.19	0.07	1.0
9	1%	2.26	3.10	0.14	0.8

standard errors displayed in the fifth column). Finally, Readers appeared to use the various rating categories in a consistent fashion—the unweighted mean-square fit indices for the rating categories are all close to the expected value of 1, as shown in the sixth column of Table 4.[8]

The third column of Figure 1 ("Students") displays the location of each student on the logit scale. Higher-achieving students appear at the top of the column, while lower-achieving students appear at the bottom. Each asterisk in this column represents 457 students, and each period represents from 1 to 456 students. This portion of the figure makes it clear that the students exhibited a wide range of levels of achievement, with a high concentration situated between the logit values of –2 and +3 logits. The mean and standard deviation of the student achievement measures equal 0.00 and 1.65, respectively. In addition, the distribution is multimodal—a common occurrence when Readers use a limited number of rating categories. For each student, there are only nine possible scores—the number of rating scale points.

The fourth column of Figure 1 ("Benchmarks") identifies the location of the 28 benchmark student essays—note that these students' essays are represented in both the third and the fourth columns. The benchmark essays were fairly evenly spread across the range of the student achievement distribution. However, there were few benchmarks located in the extremes of the distribution.

The fifth column of Figure 1 ("Time Periods") displays the location of each time period on the logit scale (We analyzed rating data obtained over four days of the AP Reading. There were two time periods each day—morning and afternoon. Time 1 [shown as "1" in the figure] refers to the average rating of all essays rated during the morning of Day 1; Time 2 [shown as "2" in the figure] refers to the average rating of all essays rated during the afternoon of Day 1, and so on.) Note that the range of the average essay rating across time periods is compact, indicating that the mean ratings do not vary much from one time period to the next. For example, Time Periods 3, 4, and 6 appear on a single line in Figure 1; the average essay rating for these three time periods ranged from 0.38 to 0.54 logits. Similarly, Time Periods 2, 5, 7, and 8 appear on a line; the average essay rating for these four time periods ranged from 0.31 to 0.37 logits. From these results, it appears that Readers, as a group, became slightly more lenient as the AP Reading progressed, since the logit measures seem to decrease over time.

This point is more clearly made in Table 5, which presents the average essay rating for each of the eight time periods (in logits), their standard errors, the "fair averages," the average of the raw ratings, and unweighted mean-square fit indices for the time periods. First, notice that the average essay rating (in logits) decreased over the course of the AP Reading, from an average of 0.49 logits for the first two days of the Reading (i.e., Times 1–4) to an average of 0.38 logits for the third and fourth days of the Reading (i.e., Times 5–8). The fourth column of the table displays the "fair average" for each of these logit measures; that is, the average essay rating at each time period, correcting for variation among the severities of Readers who rated during that particular time period. (The values reported in this column are on the 9-point AP scale.) When we examine the values in this column, we see that the average essay rating increased about 0.28 points on the 9-point rating scale over the course of the AP Reading. Second, notice that the average essay ratings were more stable during the third and fourth days of the AP Reading than during the first two days (i.e., the range of the logit measures for the first two days (i.e., Times 1–4) is 0.28 logits while the range of the logit measures for the third and fourth days (i.e., Times 5–8) is only 0.07 logits. Hence, it seems that Readers also became more consistent in the levels of severity they exercised over time. Third, note that there is not a one-to-one correspondence between the expected ratings (i.e., the "fair averages") and the averages of the raw ratings. For example, at Time 4 the average of the raw ratings is more than one rating scale point higher than the corresponding fair average; by contrast, at Time 7 the average of the raw ratings is about one-half of a rating scale point lower than the corresponding fair average. Fourth, notice that the unweighted mean-square fit values for the time periods tend to decrease over time. This indicates that, as the Reading progressed, the Readers were less likely to assign ratings that were inconsistent with those that the MFRM predicted. This may also mean that Readers became more accurate over time or, possibly, that Readers, as a group, may have used fewer rating categories as the Reading progressed.

[8] The expected value according to the model is 1, implying that the ratings, on average, are each contributing one unit of statistical information to the measurement process, as they are intended to do.

Table 5

Time Period Facet Summary for the Time Facet Model

Time Period	*Average Essay Rating (in logits)*	*SE*	*Fair Average*	*Average of the Raw Ratings*	$MS_{unweighted}$
1	0.65	0.04	4.01	4.0	1.22
2	0.37	0.04	4.24	4.9	1.00
3	0.54	0.04	4.10	4.4	1.00
4	0.40	0.04	4.22	5.3	0.94
5	0.33	0.04	4.28	4.5	1.05
6	0.38	0.04	4.24	4.5	1.04
7	0.33	0.04	4.28	3.8	0.85
8	0.31	0.04	4.29	4.1	0.94

The sixth column of Figure 1 ("Readers") displays the location of the Reader severity/leniency measures on the logit scale. Note that the range of Reader severity measures is wider than the range of the average essay ratings, indicating that Readers differed more in their individual levels of severity than the group of Readers changed in terms of their average ratings across time (i.e., compare the distributions of measures shown in the fifth and sixth columns).

Table 6 presents summary statistics for the Readers. The dispersion of Reader severity/leniency measures (in logits) is narrower than the dispersion of the student achievement measures. Specifically, the standard deviation of the Reader severity measures is 0.39, while the standard deviation of the student achievement measures is 1.65. This is encouraging, because studies back to Edgeworth (1890) have reported Reader dispersions on the order of half that of the students. The maximum and minimum values of the Reader severity measures span about two logits, which translates to about 1.8 points on the 9-point rating scale. The fair average reported in Table 6 is the average rating for each Reader once the computer program adjusts that average for the deviation of the essays in each Reader's sample from the overall average across all Readers and essays. By comparing the Readers' "fair averages," we can pinpoint those Readers who tended to use the 9-point rating scale in a different manner than other Readers (i.e., who assigned ratings that were on average lower [or higher] than those that the other Readers assigned, even after the particular essays that those Readers evaluated were taken into account). The minimum Reader fair average was 3.37, while the maximum Reader fair average was 5.18. These findings imply that, in this particular AP rating context, the difference between the most lenient and the most severe Readers is almost two rating scale points, which is surely likely to be important for credit/no credit decision making. Reader severity is precisely measured—the standard errors of the severity measures average about 0.14 logits. Readers vary in the level of consistency they exhibit. Some were more consistent in their application of the scoring guidelines than others. Specifically, the unweighted mean-square fit indices for the Readers range from 0.51 to 3.00.

Table 6

Reader Facet Summary for the Time Facet Model

	Logit	*SE*	*Fair Average*	$MS_{unweighted}$
Mean	0.00	0.14	4.22	1.01
SD	0.39	0.01	0.34	0.40
Minimum	-0.99	0.13	3.37	0.51
Maximum	1.05	0.15	5.18	3.00

Changes in Leniency/Severity (Research Questions 1 and 2)

Research Question 1 asks whether AP Readers varied in the levels of severity and leniency they exercised as the Reading progressed. Recall that the summary of the time period facet from the Time Facet model displayed in Table 5 suggests that Readers appeared to become slightly more lenient over time. To force the Reader facet of the Time Facet model to depict these differences, we scaled the data to the Interaction model (Equation [3]), anchoring each time period's estimate at the value of zero. This caused any observed differences between the average essay rating for each time period to manifest themselves in the Reader severity/leniency measures. Table 7 summarizes three indices that these analyses produced. Specifically, the second through the seventh columns display the indices that compare the ratings assigned during Times 2 through 7 with the ratings assigned during Time 1 (baseline). The first row of indices provides the average reader-by-time period index for each of the comparisons (I_{rc}). The fact that all the logit values in this row are positive jibes with the summary provided in Table 5. That is, I_{rc} is positive when a Reader assigns higher ratings during the comparison time (t) than during Time 1. Hence, the positive logit values in the first row of Table 7 indicate that, relative to their ratings at Time 1, the Readers as a group tended to become more lenient as the Reading progressed. In addition, and also consistent with the information displayed in Table 5, the amount of group-level change in the level of leniency exercised is fairly stable from one time period to the next—that is, the Readers (as a group) became more lenient from Time 1 to Time 2, but the mean logit increase in leniency did not vary appreciably across the remaining time periods.[9]

[9] The results reported in Tables 5 and 7 are consistent. For example, note that in Table 5 when we compared Time 1 (0.65) to Time 2 (0.37), the difference was 0.28, which is very close to the value of 0.26 reported in Table 7 for the comparison of Time 1 to Time 2.

Table 7

Reader-by-Time Interaction Summary for the Interaction Model

	Time-Period Severity Comparison						
	1 vs. 2	*1 vs. 3*	*1 vs. 4*	*1 vs. 5*	*1 vs. 6*	*1 vs. 7*	*1 vs. 8*
Mean I_{rc} logit increase in leniency	0.26	0.14	0.25	0.31	0.23	0.22	0.26
% $\alpha_{\lvert t_{I_{rc}} \rvert} < 0.05$	17	27	21	21	23	16	22
% $\lvert I_{rc} \rvert > 0.50$	60	62	62	65	58	64	66

The second row of Table 7 shows the percentage of Readers who exhibited a statistically significant shift in their levels of severity across time periods. The first row displays the logit difference between each time period and the first time period, which can be obtained from Table 5. The second row shows that the values that are reported here are I_{rc} absolute values that deviate from zero by a statistically significant degree. For example, as shown in the second column, 17 percent of the Readers exhibited a statistically significant shift in their level of severity or leniency between Time 1 and Time 2. Overall, the average percent of change is about 21 percent, with no evidence of any appreciable increase in the percentages across the eight time periods. That is, most of the shift toward increased leniency occurred between Time 1 and Time 3, with the flag rate becoming more stable after Time 3.

The bottom row of Table 7 reports the percentage of Readers who exhibited a change in level of severity or leniency that is large enough to be substantively meaningful at each time period relative to Time 1. That is, we judged an effect to be substantively meaningful if the absolute value of I_{rc} was greater than 0.50. As displayed in Table 7, a large percentage of Readers satisfied this criterion. That means that the statistically significant differences observed in the second row of indices are not simply due to a greater level of precision (i.e., a large number of ratings per reader)—rather, those differences are large enough to raise concern.

Another way of depicting changes in Reader severity/leniency over time is to consider the Reader severity measures obtained from the Separate model (Equation [4]). Table 8 displays descriptive statistics for the SAI_{rc} index (all subsequent time periods are compared to Time 1) generated from the Separate model. As shown in the first four rows of Table 8, the SAI_{rc} indices tended to be negative across time periods (i.e., there seemed to be an overall trend toward greater leniency over time), but that effect was most apparent between Time 1 and Time 2. That is, Readers were more lenient during Time 2 than during Time 1, but the amount of group-level change tended to level off after Time 2. It is important to note that the trend toward increasing leniency over time was not constant. In fact, there were decreases in the SAI_{rc} index between Times 2 and 3, and between Times 5 and 6, as shown in Figure 2. This figure reiterates the trend displayed in Table 8—average rater severity increased between 0.17 and 0.48 logits over time.

Table 8

SAI_{rc} Descriptive Statistics for the Separate Model

	$\mathbf{SAI}_{rc}$ *Time-Period Severity Comparison*						
	1 vs. 2	*1 vs. 3*	*1 vs. 4*	*1 vs. 5*	*1 vs. 6*	*1 vs. 7*	*1 vs. 8*
Mean	−0.40	−0.17	−0.36	−0.45	−0.40	−0.45	−0.48
SD	1.09	1.26	1.17	1.22	1.28	1.17	1.27
Minimum	−3.09	−2.67	−3.26	−3.18	−5.56	−3.43	−3.82
Maximum	1.99	3.02	2.72	2.26	2.39	2.01	2.36
% Becoming Lenient ES < −0.50	50	39	42	45	42	51	44
% Becoming Severe ES > 0.50	23	29	23	25	21	24	26

Note: ES = Effect Size.

Table 8 also depicts how pervasive the trends toward increasing severity and leniency were. The minimum and maximum values of the SAI_{rc} index are far from the null value of zero. Across time periods, the typical maximum or minimum change in Reader severity/leniency was between two and three logits, which corresponds, according to Figure 1, to about three rating scale points. When compared to the definition of a meaningfully large effect size (0.50), a large proportion of the Readers exhibited severity/leniency DRIFT effects over time. As shown in the final two rows of Table 8, on average, approximately 45 percent of the Readers were flagged because they became more lenient over time based on their SAI_{rc} values, while, on average, about 24 percent of the Readers were flagged because they became more severe over time.

Table 9 displays descriptive statistics for the $Z_{SAI_{rc}}$ associated with the indices displayed in Table 8. The $Z_{SAI_{rc}}$ statistics extend the information provided in Table 8, indicating the degree to which the SAI_{rc} change is greater than the amount expected by chance. We employed a 5 percent Type I error rate for the $Z_{SAI_{rc}}$ index (i.e., we used an absolute critical value of 2.45 to create $Z_{SAI_{rc}}$ flags). As was true for Table 8, the indices in Table 9 indicate that Readers were more likely to become more lenient over time than to become more severe (as shown by the negative values of the means displayed in the first row of the table, and by the larger percentage of Readers flagged for becoming more lenient). In addition, it is clear that

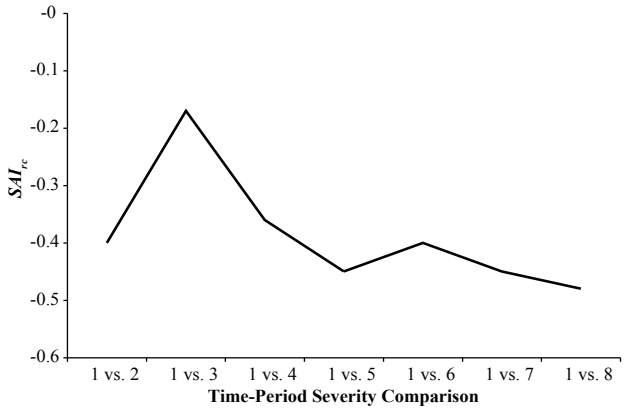


Figure 2. Separate model SAI_{rc} trend across time periods.

most of the change occurred during the first two time periods, and the amount of group-level change tended to level off after Time 2. However, these results indicate that the severity/leniency DRIFT rates are not as pervasive as the SAI_{rc} (effect size) index (Table 8) suggests. Specifically, on average, about 16 percent of the Readers exhibited a statistically significant shift toward greater leniency over time using the $Z_{SAI_{rc}}$ index, while on average less than 5 percent of the Readers exhibited a statistically significant shift toward greater severity over time using this index. The difference between the flag rates for the SAI_{rc} and $Z_{SAI_{rc}}$ indices indicates that the seven benchmark essays used to link each pair of time periods did not provide us with a great deal of parameter estimation precision. Finally, note that some Readers had quite extreme $Z_{SAI_{rc}}$ indices—the average minimum value across time periods was –5.33, and the average maximum value was 3.72. (These indices have the form of unit normal deviates.)

Beyond these group-level effects, it is informative to examine the rating patterns of individual Readers exhibiting substantively interesting cases of severity/leniency DRIFT. It is difficult to define DRIFT in an objective manner because the sampling distributions for the various DRIFT indices are unknown. One way that we evaluated DRIFT for individual Readers was to examine the average and variance of the $Z_{SAI_{rc}}$ for each Reader that we obtained from the Separate model. We reasoned that Readers having large and consistent absolute average values of $Z_{SAI_{rc}}$ would exhibit either abrupt and dramatic (or progressive and consistent) changes in their leniency/severity over time. Table 10 illustrates this point. That table contains values of SAI_{rc} and the associated value for $Z_{SAI_{rc}}$ for hypothetical Readers across seven time-period comparisons. The last two rows of Table 10 present the mean and variance of the seven values of SAI_{rc} and six values of $Z_{SAI_{rc}}$ for each Reader across time periods. Reader A exhibits a marked shift in leniency between Times 1 and 2, but then continues to exercise this new level of leniency across the remaining time periods. As

Table 9

$Z_{SAI_{rc}}$ Descriptive Statistics for the Separate Model

	$Z_{SAI_{rc}}$ *Time-Period Severity Comparison*						
	1 vs. 2	***1 vs. 3***	***1 vs. 4***	***1 vs. 5***	***1 vs. 6***	***1 vs. 7***	***1 vs. 8***
Mean	–0.62	–0.25	–0.55	–0.69	–0.59	–0.68	–0.72
SD	1.70	1.93	1.84	1.87	1.95	1.75	1.91
Minimum	–4.78	–4.10	–4.82	–4.75	–8.10	–5.15	–5.60
Maximum	3.39	4.53	4.22	3.63	3.79	2.89	3.61
% Becoming Lenient $p < 0.025$	12	15	18	18	17	14	21
% Becoming Severe $p < 0.025$	3	11	6	3	5	3	2

a result, the mean absolute value of $Z_{SAI_{rc}}$ across time is relatively large, but the variance of $Z_{SAI_{rc}}$ across time is relatively small. Reader B exhibits a gradual but consistent shift in leniency between Times 1 and 7. As a result, the mean absolute value and the variance of $Z_{SAI_{rc}}$ across time periods are moderately large. These are two cases that

Table 10

Hypothetical Trends in SAI_{rc} and $Z_{SAI_{rc}}$ for the Separate Model

Time Period	*Hypothetical Reader*			
	A	***B***	***C***	***D***
1	2.00	2.00	2.00	0.00
2	–2.00 (–6.29)	1.30 (–1.10)	–2.00 (–6.29)	0.00 (0.00)
3	–2.00 (–6.29)	0.67 (–2.09)	2.00 (0.00)	0.00 (0.00)
4	–2.00 (–6.29)	0.00 (–3.14)	–2.00 (–6.29)	0.00 (0.00)
5	–2.00 (–6.29)	–0.67 (–4.20)	2.00 (0.00)	0.00 (0.00)
6	–2.00 (–6.29)	–1.30 (–5.19)	–2.00 (–6.29)	0.00 (0.00)
7	–2.00 (–6.29)	–2.00 (–6.29)	2.00 (0.00)	0.00 (0.00)
Mean	–1.43 (–6.29)	0.00 (–3.67)	0.29 (–3.14)	0.00 (0.00)
Variance	2.29 (0.00)	2.05 (3.76)	4.57 (11.85)	0.00 (0.00)

Note: Values in each cell include the SAI_{rc} and ($z_{SAI_{rc}}$) for each hypothetical Reader. Reader A exhibits a sudden shift in leniency and then continues to exercise that new level of leniency over time. Reader B exhibits a gradual and consistent shift in leniency over time. Reader C exhibits erratic shifts in leniency/severity over time. Reader D exhibits consistent levels of leniency/severity over time.

we would be particularly interested in detecting because they constitute substantively meaningful examples of Reader severity/leniency DRIFT. On the other hand, Reader C exhibits erratic shifts in leniency and severity across time. This pattern results in a moderately large mean absolute value of $Z_{SAI_{rc}}$ across time periods and a very large value of the variance of $Z_{SAI_{rc}}$. Finally, Reader D could be characterized as an "ideal Reader" because that Reader exhibits no change in leniency/severity across time periods—behavior that results in a mean value and a variance of $Z_{SAI_{rc}}$ that are close to zero.

Because the sampling distribution of the mean of $Z_{SAI_{rc}}$ is unknown, we chose to utilize resampling procedures (Good, 1999) to estimate the shape of that null distribution. That is, we randomly sampled seven values of $Z_{SAI_{rc}}$ from the Readers included in our study, ignoring the Reader and time periods associated with each value of $Z_{SAI_{rc}}$. As a result, we could then compare the value of any Reader's comparison of Time 2 with Time 1 with the value of any other Reader's comparison of Time 3 with Time 1, and so forth. We chose 100,000 of these samples, and we then computed the mean of the $Z_{SAI_{rc}}$ statistics for each sample. The distribution of these means served as the null sampling distribution for identifying outliers among the means of the $Z_{SAI_{rc}}$ statistics for the 101 Readers in our study. Table 11 shows the average, standard deviation, and two-tailed $\alpha = 0.05$ critical values identified in this resampling procedure for the mean of $Z_{SAI_{rc}}$. The bottom two rows of this table indicate that reasonable two-tailed critical values for the mean of $Z_{SAI_{rc}}$ are –1.96 and 0.77.

Based on these values, we classified all Readers into one of three levels of differential severity/leniency: (a) *differential severity* (those with an average $Z_{SAI_{rc}}$ greater than 0.77), (b) *differential leniency* (those with an average $Z_{SAI_{rc}}$ less than –1.96), and (c) *no differential severity/leniency* (those with an average $Z_{SAI_{rc}}$ between the values of –1.96 and 0.77). We further subdivided these groups into those exhibiting *erratic* rating behavior, and those exhibiting *stable* rating behavior, based on the variability within the Readers' patterns of $Z_{SAI_{rc}}$ statistics. By our definition, erratic rating behavior occurs when a Reader's level of severity or leniency differs dramatically and in a seemingly random manner from one time period to the next. On the other hand, stable rating behavior occurs when a Reader's level of severity or leniency either remains constant or systematically increases or decreases over time. Because of the dependence between the $Z_{SAI_{rc}}$ statistics within a Reader, we could not identify a suitable resampling procedure for estimating the shape of the sampling distribution for the variance of $Z_{SAI_{rc}}$. Hence, we chose the somewhat arbitrary value of 2.22—one standard deviation above the mean of the variance of Readers' $Z_{SAI_{rc}}$ statistics—as the cutoff for classification into erratic (i.e., more extreme than this value) or stable (i.e., less than this value) examples of each differential severity/leniency category.

Table 11

Resampling Descriptive Statistics of the Mean $Z_{SAI_{rc}}$ Index

Statistic	*Mean* $Z_{SAI_{rc}}$
Mean	–0.58
SD	0.70
Minimum	–3.75
Maximum	2.37
2.5th Percentile	–1.96
97.5th Percentile	0.77

Table 12 reports the percentage of Readers in a three-by-two classification based on the mean and variance of their $Z_{SAI_{rc}}$ statistics across time periods. About 60 percent of the Readers showed little change in the values of their $Z_{SAI_{rc}}$ statistics over time (i.e., the mean $Z_{SAI_{rc}}$ for these Readers was greater than –1.96 and less than 0.77). On the other hand, about 20 percent of the Readers became differentially lenient over time (i.e., the mean value of their $Z_{SAI_{rc}}$ statistics was less than –1.96), while about 20 percent of the Readers became differentially severe over time (i.e., the mean value of their $Z_{SAI_{rc}}$ statistics was greater than 0.77). Of the Readers who displayed these tendencies, the majority of them exhibited stable rating behavior as we defined it (i.e., the variance of their $Z_{SAI_{rc}}$ statistics was close to 0.00 over time). Specifically, for 85 percent of the Readers, the variance of their $Z_{SAI_{rc}}$ statistics over time was less than 2.22. By contrast, 16 percent of the Readers exhibited erratic rating behavior (i.e., the variance of their $Z_{SAI_{rc}}$ statistics over time was greater than 2.22).

Table 13 presents specific examples of rating behaviors that a sample of Readers from each cell of Table 12 exhibited. Specifically, this table shows the time period–specific Reader severity measures (λ_{rt}) taken from the Separate model (Equation [4]) for six Readers, and the expected rating that each Reader would assign at each time period to a student whose essay was average in terms of achievement (i.e., $\theta_n = 0$).

Table 12

$Z_{SAI_{rc}}$ Index Classifications Across Time Periods

	Variance $Z_{SAI_{rc}}$		
Mean $Z_{SAI_{rc}}$	*Stable*	*Erratic*	*Total*
Differential Leniency	18%	3%	21%
None	52%	8%	60%
Differential Severity	15%	5%	20%
Total	85%	16%	101%

Note: The values indicate the percentage of Readers in each classification. The total is greater than 100 percent due to rounding.

These expected ratings (E_{nrt}) represent the model-based predicted ratings that a given Reader would assign to a single student's essay at the various time periods. If the Reader exhibits no changes in severity or leniency across time, both sets of numbers—the time period–specific severity measures and the expected ratings—should fluctuate little across time periods. For example, the third Reader (NCS—No Change Stable) is one of the 52 Readers appearing in the "None-Stable" cell of Table 12. This Reader exhibited only small changes in his or her time period–specific severity measures, ranging from a high of 1.10 logits at Time 5 to a low of 0.30 logits at Time 2. This translates to approximately one-half of a point on the 9-point rating scale.

Other Readers also exhibited little evidence of consistent change in severity/leniency over time, even though the values of their time period–specific severity measures fluctuated considerably. For example, the severity measures of Reader NCE (No Change Erratic) in Table 13 exhibited considerably greater variability than those of Reader NCS, ranging from a high of 1.78 logits at Time 5 to a low of –1.08 logits at Time 8. However, an examination of the expected ratings for Reader NCE does not reveal a tendency to become more severe (or lenient) as time progressed. Rather, this Reader's expected ratings vary considerably across time periods around a mean value of 4.23—the high is 5.46 at Time 8, and the low is 3.52 at Time 5.

We observed similar patterns for Readers exhibiting both stable and erratic behaviors. For example, the average severity measure for Reader LS (Leniency Stable) drops from 0.20 logits to –0.66 logits between the first four and the final four time periods. This translates to an increase of just over one-half of a point on the rating scale, with fairly small variability from one time period to another. On the other hand, Reader LE (Leniency Erratic) exhibits an average increase in expected ratings from the first four to the final four time periods of almost one-and-one-half points, but this Reader's variability from one time period to the next is considerable. In the case of DRIFT toward severity, Reader SS (Severe Stable) exhibits a one-point drop in expected ratings between the first and last time periods, even though the average absolute change in ratings is only 0.36 rating scale points. By contrast, Reader SE (Severe Erratic) exhibits a drop of only one-half of a rating scale point between the first and final time periods, but this Reader's average absolute change in ratings across time periods is 1.05 rating scale points. Figure 3 displays the expected ratings for all six of these Readers.

Research Question 2 focuses on the comparability of multiple indices used to flag Readers for differential severity and leniency. In our discussion of changes in Reader severity and leniency, we presented two relevant indices: (a) the reader-by-time period interaction term from the Interaction model (I_{rc}) and its associated t statistic ($t_{I_{rc}}$), and (b) the signed area index (SAI_{rc}) and its associated standardized index ($Z_{SAI_{rc}}$) from the Separate model. We also previously described how differential severity/leniency may restrict the variability of the assigned ratings, thus influencing the value of $F_{V(X)}$.

To compare these multiple indices, we examined the correlations between them. The SAI_{rc} and I_{rc} indices and the $Z_{SAI_{rc}}$ and $t_{I_{rc}}$ indices were perfectly correlated at each time period. On the other hand, $F_{V(X)}$ had a near zero correlation with the absolute values of $t_{I_{rc}}$ and $Z_{SAI_{rc}}$ ($r = -0.01$ for both).

In addition, the flag rates for statistically significant changes in Readers' levels of severity or leniency were very high when we compared the values of their $Z_{SAI_{rc}}$ and $t_{I_{rc}}$ indices. Specifically, the average rate of agreement between these two indices was 97 percent

Table 13

Differential Severity/Leniency Examples from the Separate Model

			Time Period							
Reader	***Rating Behavior***	***Index***	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***
LS	Leniency Stable	λ_{rt}	2.07	–1.02	0.05	–0.32	–0.62	–0.14	–1.36	–0.52
		E_{nrt}	3.35	5.41	4.56	4.83	5.06	4.69	5.73	4.98
LE	Leniency Erratic	λ_{rt}	2.57	–0.04	0.05	1.74	–0.07	–2.99	–0.61	0.07
		E_{nrt}	3.04	4.62	4.56	3.54	4.64	7.32	5.05	4.55
NCS	No Change Stable	λ_{rt}	0.71	0.30	1.06	0.38	1.10	0.81	0.79	0.96
		E_{nrt}	4.14	4.40	3.94	4.34	3.91	4.08	4.09	3.99
NCE	No Change Erratic	λ_{rt}	0.92	0.13	1.27	1.33	1.78	1.45	–0.61	–1.08
		E_{nrt}	4.01	4.51	3.81	3.78	3.52	3.71	5.05	5.46
SS	Severe Stable	λ_{rt}	–0.81	0.47	0.64	0.20	–0.25	0.81	0.57	0.72
		E_{nrt}	5.22	4.29	4.18	4.46	4.77	4.08	4.22	4.13
SE	Severe Erratic	λ_{rt}	–1.16	–0.04	1.06	–0.15	0.89	1.23	–2.13	–0.52
		E_{nrt}	5.54	4.62	3.94	4.70	4.03	3.84	6.49	4.98

Note: The values of λ_{rt} are reported on a logit scale. The values of E_{nrt} are reported on the 9-point rating scale.

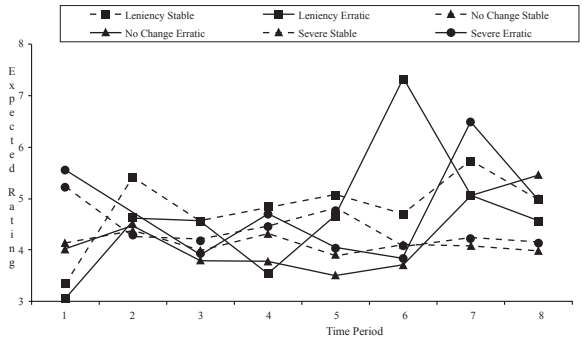


Figure 3. Changes in severity/leniency examples.

across the seven time-period comparisons. Coefficient kappa, a measure of the degree to which the observed agreement exceeds that attainable by chance, was also very high—the average value across the seven time-period comparisons was 0.91.

Similarly, the agreement between flags for individual Readers was very high for the $t_{I_{rc}}$ and the $Z_{SAI_{rc}}$ indices, but was fairly low between these indices and the $F_{V(X)}$ index. Specifically, 79 percent of the Readers were identically categorized as exhibiting differential severity, differential leniency, or no change between Times 1 and 2 by the $Z_{SAI_{rc}}$ and $t_{I_{rc}}$ indices, and 24 percent of the Readers were flagged for differential severity or leniency by both of these indices. On the other hand, the $Z_{SAI_{rc}}$ and $F_{V(X)}$ indices categorized only 68 percent of the Readers identically between Times 1 and 2, and only 2 percent of the Readers were jointly flagged by both of these indices.

Changes in Category Use and Accuracy (Research Questions 3, 4, 5, and 6)

In this section, we discuss evidence relevant to the research questions focusing on differential scale category use and differential accuracy/inaccuracy over time. That is, we focus on whether Readers exhibit differential scale category use (Research Question 3), whether Readers exhibit differential accuracy/inaccuracy (Research Question 5), and the degree to which different indices relevant to detecting these trends produce comparable results (Research Questions 4 and 6). We summarize the following indices here because, as described in Table 1, they are all jointly relevant to the detection of differential scale category use and differential accuracy/inaccuracy effects: $Z_{E-R_c,E-R_b}$ and $Z_{SR-ROR_c,SR-ROR_b}$ (of primary interest for both differential scale category use and differential accuracy/inaccuracy), and $F_{V(X)}$ and F_{fit} (of secondary interest for both differential scale category use and differential accuracy/inaccuracy). Specifically, (refer to Table 1) differential scale category use produces $Z_{E-R_c,E-R_b}$ values that are less than 0.00, $F_{V(X)}$ values that are less than 1, and F_{fit} values that are less than 1. By contrast, differential accuracy and inaccuracy effects produce $Z_{SR-ROR_c,SR-ROR_b}$ values that are positive and negative, respectively. Increases in accuracy also result in F_{fit} values that are less than 1, while decreases in accuracy result in values of F_{fit} that are greater than 1.

The first set of questions we consider focuses on the existence of differential scale category use and differential accuracy/inaccuracy (Research Questions 3 and 5). Examination of the rating scale category use across all time periods reveals that there is good reason to be concerned about differential scale category use. Specifically, only 40 percent of the Readers utilized all

Table 14

Descriptive Statistics for Static Central Tendency and Accuracy Indices

		Time Period							
Index	***Name***	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***
$MS_{weighted}$	Weighted Reader Mean-Square Fit Index	1.13 (0.82)	1.02 (0.69)	0.98 (0.66)	0.98 (0.78)	1.08 (0.71)	1.09 (0.79)	0.80 (0.58)	0.81 (0.59)
$MS_{unweighted}$	Unweighted Reader Mean-Square Fit Index	1.13 (0.80)	1.04 (0.75)	0.96 (0.64)	0.96 (0.74)	1.15 (0.84)	1.09 (0.81)	0.82 (0.59)	0.85 (0.75)
$r_{res,exp}$	Correlation of Reader Residuals and Expected Ratings	0.03 (0.17)	0.02 (0.22)	0.05 (0.25)	0.02 (0.17)	0.01 (0.19)	0.01 (0.15)	0.03 (0.14)	0.06 (0.16)
r_{SR-ROR}	Single Reader–Rest of the Readers Correlation	0.78 (0.17)	0.82 (0.13)	0.91 (0.07)	0.80 (0.18)	0.89 (0.11)	0.79 (0.17)	0.85 (0.11)	0.91 (0.07)
S_X	Standard Deviation of Raw Ratings	1.46 (0.24)	1.42 (0.27)	1.40 (0.24)	1.39 (0.25)	1.47 (0.23)	1.38 (0.24)	1.40 (0.21)	1.46 (0.22)

Note: Values in each cell include the mean of the index in question (and its standard deviation).

nine rating categories during the entire Reading. A small portion of the Readers (i.e., four of them—4 percent) utilized as few as six categories, while a sizeable portion of the Readers utilized only seven or eight rating categories (i.e., 25 percent and 32 percent, respectively).

Table 14 displays the mean and standard deviation of each index relevant to the detection of static central tendency and accuracy/inaccuracy that we generated using the Separate model at each time period. These indices are useful for detecting pervasive trends in Reader performance over time. The first two rows of Table 14 indicate that overall Reader fit improved as the Reading progressed (i.e., the means of the Reader mean-square fit indices, $MS_{weighted}$ and $MS_{unweighted}$, generally tended to decrease over time). In addition, the means of the single Reader–rest of the Readers correlation coefficients (r_{SR-ROR}) showed a very slight increase over time. There was little or no change in the means of the correlations of Reader residuals and expected ratings ($r_{res,exp}$) and the means of the standard deviations of the raw ratings (S_X). We also note that standard deviations of the indices shown in Table 14 varied slightly over time, but none displayed a pervasive trend. We caution against drawing inferences about trends in DRIFT based on any of the indices shown in Table 14 because their interpretations may not be intuitively apparent. Rather, we will examine specific, more meaningful indicators of DRIFT when we discuss the results presented in Table 15.

Table 15 displays the mean and standard deviation of each statistic relevant to the detection of differential scale category use and differential accuracy/inaccuracy effects for each time period compared to the first time period (the initial time period benchmark) using the Separate model. The index of primary interest for detecting differential scale category use, $Z_{E-R_c,E-R_b}$, does not vary much across time—an indicator that Readers as a group do not exhibit evidence of differential scale category use. This is supported by the nearly constant mean values of $F_{V(X)}$, the index of secondary interest for detecting differential scale category use.

By contrast, the mean values of $Z_{SR-ROR_c,SR-ROR_b}$ provide weak evidence that Readers tended to become more accurate as the Reading progressed. We see that Readers became more accurate from Time 1 to Time 3 ($Z_{SR-ROR_c,SR-ROR_b}$ = 0.74). However, their accuracy appears to have waned in the middle of the Reading (e.g., comparing mean values for Time 1 to Time 4, $Z_{SR-ROR_c,SR-ROR_b}$ = 0.14, and Time 1 to Time 6, $Z_{SR-ROR_c,SR-ROR_b}$ = 0.05) and then increased again toward the end of the Reading (e.g., comparing mean values for Time 1 to Time 7, $Z_{SR-ROR_c,SR-ROR_b}$ = 0.31 and Time 1 to Time 8, $Z_{SR-ROR_c,SR-ROR_b}$ = 0.67). Although the mean values of $Z_{SR-ROR_c,SR-ROR_b}$ vary quite a bit across the time-period comparisons, these values tend to get larger over time—an indicator of generally increasing accuracy as the Reading progressed.

The positive mean values of the F_{fit} indices support this notion. Specifically, the average Reader misfit increased slightly when we compare Time 1 to Times 2 through 6. However, the mean values of the F_{fit} indices were close to 1 for the comparisons of Time 1 to Times 7 and 8, suggesting that this increase in misfit was only temporary. As we discussed in our summary of Table 1, these changes in the mean F_{fit} indices over time could have two possible interpretations: (a) they might suggest that during the beginning and middle of the Reading, the Readers were more inaccurate than they were at the end of the Reading, or (b) they could suggest that during the beginning and middle of the Reading, the Readers used the central rating scale categories more frequently than they did at the end of the Reading—an interpretation that our examination of the $Z_{E-R_c,E-R_b}$ indices do not support.

Table 16 shows the percentage of individual Readers exhibiting statistically significant changes over time in their single Reader–rest of the Readers correlations, the correlations of Reader residuals and expected ratings, the Reader mean-square fit indices, and the standard

Table 15

Descriptive Statistics for DRIFT Central Tendency and Accuracy Indices

	Time-Period Comparison						
Statistic	***1 vs. 2***	***1 vs. 3***	***1 vs. 4***	***1 vs. 5***	***1 vs. 6***	***1 vs. 7***	***1 vs. 8***
$Z_{E-R_c,E-R_b}$	0.00 (1.02)	0.14 (1.38)	–0.06 (1.11)	–0.11 (1.09)	–0.06 (1.02)	0.03 (1.03)	0.14 (1.22)
$Z_{SR-ROR_c,SR-ROR_b}$	0.13 (0.81)	0.74 (0.90)	0.14 (0.88)	0.54 (0.75)	0.05 (0.80)	0.31 (0.81)	0.67 (0.78)
$F_{fit-weighted}$	1.27 (1.23)	1.38 (2.20)	1.32 (1.50)	1.42 (1.53)	1.37 (1.32)	1.05 (0.95)	1.00 (0.87)
$F_{fit-unweighted}$	1.32 (1.29)	1.39 (2.55)	1.30 (1.55)	1.61 (2.52)	1.39 (1.45)	1.08 (0.99)	1.01 (0.96)
$F_{V(X)}$	1.06 (0.62)	1.00 (0.43)	1.01 (0.64)	1.10 (0.48)	0.99 (0.46)	1.00 (0.43)	1.11 (0.64)

Note: The values shown include the mean, and the (standard deviation) for each time-period comparison.

Table 16

DRIFT Flag Rates for Differential Scale Category Use and Accuracy

Statistic	*Flag*	*Time-Period Comparison*						
		1 vs. 2	*1 vs. 3*	*1 vs. 4*	*1 vs. 5*	*1 vs. 6*	*1 vs. 7*	*1 vs. 8*
$Z_{E-R_c, E-R_b}$	% Decreasing	4	6	6	6	3	3	7
	% Increasing	3	10	3	2	2	2	5
$Z_{SR-ROR_c, SR-ROR_b}$	% Decreasing	1	0	1	0	2	0	0
	% Increasing	1	8	4	3	1	3	5
$F_{fit-weighted}$	% Decreasing	2	4	6	4	6	7	7
	% Increasing	1	2	1	2	2	0	0
$F_{fit-unweighted}$	% Decreasing	2	3	7	5	6	6	6
	% Increasing	2	2	1	3	2	0	0
$F_{V(X)}$	% Decreasing	9	16	14	7	19	13	8
	% Increasing	6	3	5	4	5	5	7

Note: The values indicate the percentage of indices that meet the criteria for statistical significance for each time-period comparison.

deviations of the raw ratings. These figures differ from those displayed in Table 15 because the values shown in Table 15 are the average for all Readers included in the study. By contrast, Table 16 reports the percentage of *individual* Readers showing change over time in each of the indices (i.e., DRIFT "flag rates").

The results displayed in Table 16 reveal that, with the exception of the comparison of Times 1 and 3, the percentage of Readers showing a statistically significant decrease in the correlation of Reader residuals and expected ratings (i.e., the $Z_{E-R_c, E-R_b}$ statistic) is slightly higher than the percentage of Readers showing a statistically significant increase in this correlation. This finding suggests that, as the Reading progressed, Readers may have made greater use of the central rating scale categories (i.e., differential scale category use). A slightly higher percentage of Readers showed increases over time in their single Reader–rest of the Readers correlation than showed decreases in this statistic (see the $Z_{SR-ROR_c, SR-ROR_b}$ rates), which suggests increasing accuracy over time. Note that larger percentages of Readers showed lower levels of misfit over time than showed higher levels of misfit ($F_{fit-weighted}$ and $F_{fit-unweighted}$). This finding suggests either increasing use of central rating scale categories, or increasing accuracy. Finally, a larger percentage of Readers showed decreasing standard deviations of their raw ratings over time than showed increasing standard deviations (see the $F_{V(X)}$ flag rates). This finding supports the notion that by the middle of the Reading, the Readers were making greater use of the central scale categories than they were at the beginning. Furthermore, they continued to utilize those middle categories as the Reading progressed. In all cases, although the observed flag rates are elevated slightly above the 5 percent expected rate, they are not elevated sufficiently to conclude that Readers exhibited a pervasive trend toward DRIFT. These flag rates are also displayed in Figure 4.

Given the results displayed in Table 16, one might wonder to what degree the Readers who exhibited evidence of DRIFT showed continual change in their rating behavior over time. Table 17 displays the frequencies with which individual Readers showed statistically significant changes in different DRIFT indices across various numbers of time periods. In all cases, the majority of Readers exhibited no evidence of statistically significant DRIFT. Most commonly, Readers who exhibited evidence of statistically significant DRIFT showed change in their rating behavior for only one time-period comparison. This suggests that when Readers changed their rating behavior, they did so only temporarily. That is, they did not sustain that DRIFT behavior over time.

However, it is important to point out that there are individual Readers who showed evidence of sustained DRIFT, which resulted in marked changes in their rating behavior that they maintained across multiple time periods. For example, three Readers exhibited five statistically significant decreases in the correlation of Reader residuals and expected ratings (i.e., changes in their $Z_{E-R_c, E-R_b}$ indices)—an indicator of sustained differential scale category use. By contrast, one Reader

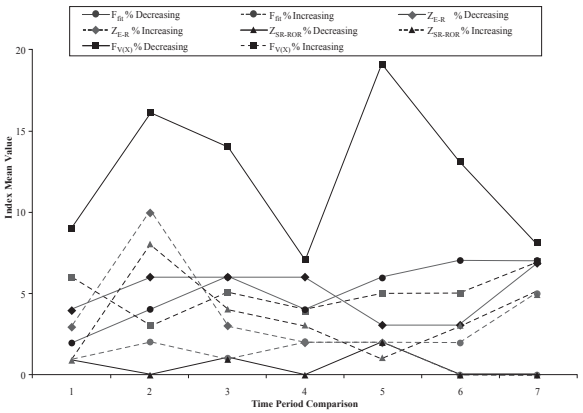


Figure 4. Trend toward differential accuracy across time periods.

Table 17

DRIFT Flag Trends for Differential Scale Category Use and Accuracy

Statistic	Flag	Number of Statistically Significant Changes						
		0	1	2	3	4	5	6
$Z_{E-R_c, E-R_b}$	% Decreasing	84	10	2	2	0	3	0
	% Increasing	85	9	4	2	1	0	0
$Z_{SR-ROR_c, SR-ROR_b}$	% Decreasing	98	2	1	0	0	0	0
	% Increasing	87	8	4	0	1	1	0
$F_{fit-weighted}$	% Decreasing	76	19	4	1	0	0	1
	% Increasing	97	2	1	0	1	0	0
$F_{fit-unweighted}$	% Decreasing	79	16	2	3	0	0	1
	% Increasing	96	2	2	0	1	0	0
$F_{V(X)}$	% Decreasing	63	17	8	6	3	1	3
	% Increasing	81	11	5	2	2	0	0

Note: The values indicate the percentage of indices that meet the criteria for statistical significance for the number of time periods shown.

exhibited four statistically significant increases in the single Reader–rest of the Readers correlation over time (i.e., changes in his or her $Z_{SR-ROR_c, SR-ROR_b}$ indices). Another Reader exhibited five statistically significant increases in this statistic. For both of these Readers, the increases in this correlation indicated that they became progressively more accurate as the Reading moved forward. Only one Reader showed six statistically significant decreases in the value of the associated *F* statistics over time (i.e., for the $F_{fit-weighted}$ values and the $F_{fit-unweighted}$ values for six of the seven comparisons with Time 1). For this Reader, the continued improvement in fit indicates that the Reader became more accurate as the Reading progressed, and that he or she was able to maintain that positive change in rating behavior over time. Finally, four Readers exhibited a statistically significant decrease in the standard deviations of their ratings across five or six of the time-period comparisons (i.e., changes in their $F_{V(X)}$ indices)—an indicator that these Readers may have continued to overuse the central rating categories for extended periods of time.

Table 18 presents actual examples of Readers in our study who displayed the patterns we identified in Table 17. The first example shows a Reader who exhibited a temporary decrease in the Reader residuals–expected ratings correlation—a pattern consistent with increased use of the central rating categories as time progresses (i.e., differential scale category use). For five of the seven comparisons with Time 1 (i.e., Time 2 through Time 6), the Reader's residuals–expected ratings correlation shows a statistically significant decrease (i.e., the value of $Z_{E-R_c, E-R_b}$ is less than –1.96). However, the Reader's residuals–expected ratings correlation returns to its near-zero value during Time 7 and Time 8. Notice that the residuals–expected ratings correlations are somewhat large—about 0.49 on average across these five time periods. The third row in this example shows the impact of this Reader's DRIFT on the ratings this Reader assigned. Specifically, when we look at how the Reader used the rating categories for those time-period comparisons in which the $Z_{E-R_c, E-R_b}$ is statistically significant, we see that over 90 percent of this Reader's ratings were 4, 5, or 6. By contrast, only 59 percent of the Reader's ratings fell into those three central categories at Time 1—the baseline. Incidentally, this Reader's increased use of the central scale categories manifests as *increases* in the values of $F_{fit-weighted}$ over time (i.e., four of the seven time-period comparisons are statistically significant)—a trend that is *opposite* what one might expect. However, other indicators (not shown in Table 18) do not look suspicious for this Reader (e.g., only one of the seven values of $Z_{SR-ROR_c, SR-ROR_b}$ is statistically significant). It is interesting to note that this shift in scale category usage initially occurred during the first day of the Reading, continued for the second and third days of the Reading (i.e., days during which Table Leaders dedicated considerable time and effort to monitoring and retraining Readers), and then began to dissipate during the fourth day of the Reading.

The second example presents a case in which the Reader's level of agreement with other Readers increased over time (i.e., the value of $Z_{SR-ROR_c, SR-ROR_b}$ increased). At Time 1, this Reader's ratings showed a very low level of agreement with the other Readers' ratings (i.e., $r_{SR-ROR} = 0.49$). The Reader's level of agreement with the other Readers increased dramatically, and the Reader continued to show high levels of agreement for six of the remaining seven time periods; note that four of these six increases in the $Z_{SR-ROR_c, SR-ROR_b}$ were statistically significant. In fact, the average value of r_{SR-ROR} over the remaining seven time periods equals 0.89. As was true for the last example, existence of this DRIFT pattern influences values of $F_{fit-weighted}$. Specifically, three of the seven $F_{fit-weighted}$ comparisons are statistically significant, indicating a decrease in Reader misfit from Time 1. Only two of the comparisons for $Z_{E-R_c, E-R_b}$ are

Table 18

Differential Scale Category Use and Accuracy/Inaccuracy Examples from the Separate Model

		Time (Comparison)							
Pattern	*Statistic*	*1*	*(1 vs. 2) 2*	*(1 vs. 3) 3*	*(1 vs. 4) 4*	*(1 vs. 5) 5*	*(1 vs. 6) 6*	*(1 vs. 7) 7*	*(1 vs. 8) 8*
$Z_{E-R_c,E-R_b}$ Decrease	$Z_{E-R_c,E-R_b}$		(–1.98)	(–3.85)	(–2.20)	(–2.46)	(–2.52)	(–1.23)	(–0.17)
	$r_{E,R}$	0.12	–0.46	–0.71	–0.42	–0.41	–0.43	–0.16	0.09
	% of 4 < X < 6	59	90	94	91	88	90	66	71
	$F_{fit-weighted}$		(2.54)	(19.15)	(11.92)	(9.62)	(7.54)	(2.00)	(1.15)
$Z_{SR-ROR_c,SR-ROR_b}$ Increase	$Z_{SR-ROR_c,SR-ROR_b}$	0.49	(1.49)	(2.20)	(1.99)	(1.59)	(0.10)	(2.49)	(1.99)
	r_{SR-ROR}		0.92	0.97	0.96	0.93	0.54	0.98	0.96
	$F_{fit-weighted}$		(0.23)	(0.18)	(0.12)	(0.21)	(0.36)	(0.07)	(0.06)
$F_{fit-weighted}$ Decrease	$F_{fit-weighted}$		(0.20)	(0.05)	(0.14)	(0.17)	(0.11)	(0.13)	(0.04)
	$MS_{weighted}$	4.38	0.88	0.20	0.60	0.73	0.48	0.58	0.19
	r_{SR-ROR}	0.86	0.91	0.97	0.85	0.93	0.92	0.97	0.98
	% of 4 < X < 6	37	65	66	71	65	65	66	74
$F_{V(X)}$ Decrease	$F_{V(X)}$	3.19	(0.99)	(0.53)	(0.50)	(0.35)	(0.59)	(0.55)	(0.56)
	$V(X)$		3.15	1.68	1.60	1.13	1.88	1.76	1.80

statistically significant (not shown in Table 18). It is likely that initially this Reader had difficulty applying the rating scale but learned to apply it in a consistent manner by the middle of the first day of the Reading. There is also something suspicious about the sudden and temporary decline in performance at Time 6—perhaps for some reason this Reader had a bad afternoon on the third day of the Reading, or perhaps fatigue or boredom set in temporarily.

The third example presents a case for which the values of $F_{fit-weighted}$ indicated a decrease in Reader fit over time. Specifically, the decrease in the value of $MS_{weighted}$ was statistically significant for six of the seven comparisons made with the Time 1 value of this index. Interestingly, there was only a small increase in the Reader's accuracy after Time 1, and the value of the Reader's r_{SR-ROR} is even lower than the value at Time 1 for one of the subsequent times (Time 4). As a result, none of the values of $Z_{SR-ROR_c,SR-ROR_b}$ is statistically significant. In addition, only one of the values of $Z_{E-R_c,E-R_b}$ is statistically significant (not shown in Table 18). As shown in the final row for this example, the decrease in fit may reflect the Reader's increased use of the central rating scale categories over time (i.e., differential scale category use). It seems that at the beginning of the Reading this Reader used a wider range of rating scale categories than later on in the Reading. At Time 1, 37 percent of the Reader's ratings were 4, 5, or 6. However, from Time 2 forward, 65 to 74 percent of the Reader's ratings fell into those three central categories.

The fourth example depicts a case for which the values of $V(X)$ change over time. In this case, following Time 2, there is a substantial and sustained decrease in the variability of this Reader's assigned ratings—all comparisons with Time 1 are statistically significant from Time 3 and beyond. This particular Reader's level of severity did not show statistically significant changes across time periods. The Reader's severity at Time 1 was –0.71 logits. For Time 2 through Time 8, the Reader's severity measures ranged from a low of –1.10 logits to a high of –0.30 logits. Similarly, the Reader's level of agreement with other Readers varies little across the time periods, ranging from a low of 0.63 to a high of 0.96 (values not shown in Table 18)—no comparisons of this index with its Time 1 value were statistically significant. Furthermore, although three of the values of $Z_{E-R_c,E-R_b}$ are statistically significant (i.e., those comparing Times 4, 5, and 8 to Time 1—again, these values are not shown in Table 18), there seems to be no direct correspondence between the values of this statistic and those of $F_{V(X)}$. Finally, there also seems to be no correspondence between this index and the values of $F_{fit-weighted}$. In this particular case, only one of the $F_{fit-weighted}$ indices is statistically significant. Other than reduced variance in the assigned ratings, we can identify no clear potential cause for the pattern of decreasing $F_{V(X)}$ values in this case.

Research Questions 4 and 6 focus on the comparability of multiple indices used to flag Readers for differential scale category use and accuracy/inaccuracy. Table 19 displays the average correlation across time periods between the two Reader mean-square fit indices ($MS_{weighted}$ and $MS_{unweighted}$), the standard deviation of the raw ratings ($SD(X)$), and the correlations between a single Reader and the rest of the Readers (r_{SR-ROR}) and between the Reader residuals and expected ratings ($r_{E,R}$). In all cases, we estimated these values using the Separate model.

First, note the very strong positive correlation (0.95) between $MS_{weighted}$ and $MS_{unweighted}$. It is likely that a Reader

Table 19

Correlations Between Static Reader Effect Indices

	Index				
Index	$r_{E,R}$	r_{SR-ROR}	$MS_{weighted}$	$MS_{unweighted}$	SD (X)
$r_{E,R}$	1.00	0.47	–0.06	–0.12	0.53
r_{SR-ROR}		1.00	–0.63	–0.63	0.10
$MS_{weighted}$			1.00	0.95	0.21
$MS_{unweighted}$				1.00	0.17
$SD(X)$					1.00

would be flagged for DRIFT similarly by these two indices. Second, note the fairly strong negative correlation (–0.63) between the Reader mean-square fit indices and the single Reader–rest of the Readers correlation (r_{SR-ROR}). Hence, it is clear that higher levels of Reader accuracy are associated with lower values of Reader misfit. Third, there is also a fairly strong positive correlation (0.53) between the Reader residuals–expected ratings and the standard deviation of the raw ratings ($SD(X)$). This relationship is consistent with our expectations—negative values of $r_{E,R}$ should indicate overuse of the central rating categories, which would also reduce the magnitude of $SD(X)$. However, we also know that changes in a Reader's leniency or severity could reduce the magnitude of $SD(X)$ (i.e., a Reader who assigns only low or high ratings will assign a distribution of tightly clustered ratings). Fourth, the moderately strong positive correlation (0.47) between r_{SR-ROR} and $r_{E,R}$ indicates that lower levels of accuracy are associated with Readers' overuse of the central rating scale categories. Fifth, the standard deviation of the raw ratings exhibits weak positive correlations (0.17 and 0.21) with the Reader mean-square fit indices—a possible indication that the mean-square fit indices are sensitive to overuse of the central rating scale categories.

Table 20 displays the average correlations between the several statistical significance indicators of differential scale category use and accuracy/inaccuracy. Predictably, the correlation between the two indices based on the Reader mean-square fit indices is very strong and positive (0.97). In addition, there are moderately strong negative correlations (–0.56 and –0.53) between these two indices and the values of $Z_{SR-ROR_c,SR-ROR_b}$. That is, lower single Reader–rest of the Readers correlations are associated with higher Reader mean-square fit indices—the lower the agreement between Readers, the greater the misfit. There is also a moderately strong positive correlation (0.61) between the values of $Z_{SR-ROR_c,SR-ROR_b}$ and $Z_{E-R_c,E-R_b}$, an indication that decreases in accuracy over time are associated with decreases in the correlation between Reader residuals and expected ratings (i.e., an increase in the use of the central rating scale categories). And, there is a predictably moderately strong relationship (0.43) between $Z_{E-R_c,E-R_b}$ and $F_{V(X)}$. As the value of $Z_{E-R_c,E-R_b}$ decreases (i.e., the use of the central rating scale categories increases), so does the value of $F_{V(X)}$ (i.e., the variance of the ratings decreases). Finally, there are also weak negative correlations (–0.20 and –0.24) between the Reader mean-square fit indices and $Z_{E-R_c,E-R_b}$. This means that increases in misfit over time may be associated with increased use of central rating categories over time.

Table 21 displays the rates with which each of these indices flagged Readers in common for differential scale category use and accuracy/inaccuracy between Times 1 and 2. Each entry in this table indicates the percentage of Readers who were identically categorized by the two indices. The percentage in parentheses indicates the conditional percentage of Readers who were flagged by the row index, given that they were flagged by the column index. The high proportion of identical categorizations is due to the fact that a large percentage of Readers were categorized as exhibiting no Reader effect (i.e., about 90 percent for each index). The noteworthy relationships occur between the two Reader mean-square fit indices (which exhibit high levels of agreement between conditional flag rates), between $F_{V(X)}$ and the two Reader mean-square fit indices (which exhibit low levels of agreement between conditional flag rates), and between $F_{V(X)}$ and $Z_{E-R_c,E-R_b}$ (which exhibit moderate levels of agreement between conditional flag rates). These figures reveal that each of the indices we chose flags a unique subset of Readers for DRIFT, and that $F_{V(X)}$ is indeed influenced by multiple types of DRIFT.

Table 20

Correlations Between Dynamic Reader Effect Indices

	Index				
Index	$Z_{E-R_c,E-R_b}$	$Z_{SR-ROR_c,SR-ROR_b}$	$F_{MS_{weighted}}$	$F_{MS_{unweighted}}$	$F_{V(X)}$
$Z_{E-R_c,E-R_b}$	1.00	0.61	–0.20	–0.24	0.43
$Z_{SR-ROR_c,SR-ROR_b}$		1.00	–0.56	–0.53	0.12
$F_{MS_{weighted}}$			1.00	0.97	–0.03
$F_{MS_{unweighted}}$				1.00	–0.06
$F_{V(X)}$					1.00

Table 21

Common Flag Rates Between Dynamic Reader Effect Indices

	Index				
Index	$Z_{E-R_c,E-R_b}$	$Z_{SR-ROR_c,SR-ROR_b}$	$F_{MS_{weighted}}$	$F_{MS_{unweighted}}$	$F_{V(X)}$
$Z_{E-R_c,E-R_b}$	—	91% (0%)	90% (0%)	89% (0%)	86% (27%)
$Z_{SR-ROR_c,SR-ROR_b}$	9%	—	95% (0%)	94% (0%)	83% (0%)
$F_{MS_{weighted}}$	10%	5%	—	99% (75%)	84% (7%)
$F_{MS_{unweighted}}$	11%	6%	1%	—	83% (7%)
$F_{V(X)}$	14%	17%	16%	17%	—

Note: The values shown in the upper off-diagonal are the percentages of Readers whom the two indices categorized identically, while the values shown in the lower off-diagonal are the percentages of Readers whom the two indices categorized differently. The values in parentheses indicate the percentage of Readers flagged by the column index that were also flagged by the row index.

Connection and Agreement/ Accuracy Comparisons (Research Questions 7 and 8)

Research Question 7 asks how DRIFT statistics from the benchmark strategy (i.e., requiring Readers to read a common set of essays) compare to those from the multiple-choice strategy (i.e., using the common multiple-choice items to which students respond to establish connections between Readers, which is current AP practice). Research Question 8 asks how DRIFT statistics from the experienced Readers strategy (i.e., an accuracy-based strategy that anchors student achievement measures on values derived from highly experienced Readers' consensus ratings of benchmark essays, which are then used in subsequent analyses to establish connections between Readers) compare to those from the benchmark and multiple-choice strategies, both of which are agreement-based strategies.

To answer these two research questions, we examined the correlations between the values of the indices from the Many-Facet Rasch Measurement analyses—$Z_{SAI_{rc}}$, $Z_{E-R_c,E-R_b}$, and $F_{MS_{weighted}}$—generated under each of these connection strategies. We did not consider $F_{MS_{unweighted}}$, since it is strongly correlated with $F_{MS_{weighted}}$, and we did not consider the raw score indices because they would not vary across connection strategies.

The correlations shown in Table 22 reveal that the two most common connection strategies, benchmark and multiple choice (Research Question 7), produce indices that starkly differ in their depictions of DRIFT. That is, when we compare the DRIFT indices produced from the benchmark and multiple-choice strategies, the Readers are rank ordered based on their values of $Z_{SAI_{rc}}$ in only a roughly similar manner. Specifically, the across-time period correlation comparisons range from a low of 0.24 to a high of 0.35 (with an average correlation of 0.30). By contrast, the rank ordering of Readers based on their

Table 22

Correlations Between DRIFT Indices for Various Connection Strategies

		Comparison						
Anchoring Comparison	**DRIFT Statistic**	**1 vs. 2**	**1 vs. 3**	**1 vs. 4**	**1 vs. 5**	**1 vs. 6**	**1 vs. 7**	**1 vs. 8**
Benchmark vs. Multiple Choice	$Z_{SAI_{rc}}$	0.24	0.29	0.34	0.28	0.35	0.28	0.30
	$Z_{E-R_c,E-R_b}$	0.22	0.21	0.05	0.12	0.12	0.15	0.13
	$F_{MS_{weighted}}$	0.27	–0.14	–0.08	0.03	–0.18	–0.02	0.11
Benchmark vs. Experienced Readers	$Z_{SAI_{rc}}$	0.98	0.99	0.97	0.98	0.99	0.99	0.98
	$Z_{E-R_c,E-R_b}$	0.36	0.35	0.50	0.31	0.32	0.47	0.51
	$F_{MS_{weighted}}$	–0.06	–0.08	0.04	–0.05	–0.02	–0.12	0.06
Experienced Readers vs. Multiple Choice	$Z_{SAI_{rc}}$	0.21	0.30	0.31	0.25	0.32	0.27	0.26
	$Z_{E-R_c,E-R_b}$	0.31	0.21	0.25	0.15	0.29	0.25	0.25
	$F_{MS_{weighted}}$	–0.12	–0.02	0.01	0.01	0.03	0.06	0.09

Note: The values indicate the percentage of indices that meet the criteria for statistical significance for each time-period comparison.

values of $Z_{E-R_c, E-R_b}$ and $F_{MS_{weighted}}$ is even less similar for the benchmark and multiple-choice strategies, with average between-strategy correlations of 0.14 and 0.00, respectively. It seems that the benchmark and multiple-choice strategies depict DRIFT in markedly different ways.

Similarly, the experienced Readers and multiple-choice strategies (one part of Research Question 8) also differ in their depictions of DRIFT. In this case, the average between-strategy correlations for the $Z_{SAI_{rc}}$, $Z_{E-R_c, E-R_b}$, and $F_{MS_{weighted}}$ indices equal 0.27, 0.24, and 0.01, respectively. Again, these correlations are far too low to consider any of these indices to be interchangeable indicators of DRIFT across different strategies for creating connectivity within rating data.

On the other hand, when we look at the correlations between some of the indices obtained from the benchmark and experienced Readers strategies, the results appear somewhat more promising. Specifically, the $Z_{SAI_{rc}}$ statistics obtained from implementation of these two connection strategies are nearly identical, showing an average correlation of 0.98. In addition, the $Z_{E-R_c, E-R_b}$ statistics demonstrate a weak positive relationship, showing an average correlation of 0.40. However, the values of $F_{MS_{weighted}}$ statistics show, on average, a near zero correlation (i.e., –0.03, to be exact).

Summary of the Results

In the paragraphs that follow, we summarize the key results we obtained from our analyses for each research question we posed.

1. *Are there any AP English Literature and Composition Readers whose levels of severity change as the Reading progresses (i.e., Readers who exhibit differential severity or differential leniency over time)? If there are such Readers, at what point in the Reading do these changes become apparent?*
    - A greater percentage of Readers in this study became somewhat more lenient over time than became more severe. Specifically, for any given time period, 12 to 21 percent of the Readers rated significantly more leniently than they did at Time 1, according to their $Z_{SAI_{rc}}$ statistics (see Table 9). Most of the shift toward increased leniency occurred between Time 1 and Time 3, with the flag rate becoming more stable after Time 3. By contrast, for any given time period, 2 to 11 percent of the Readers rated significantly more severely than they did at Time 1, according to their $Z_{SAI_{rc}}$ statistics.
    - Individually, about 18 percent of the Readers became progressively more lenient over time, while about 15 percent became progressively more severe (see Table 12). In terms of the 9-point AP scale, this change in rating behavior over time would translate to a shift of about one-half of a rating scale point in each direction. Only about 3 percent of the Readers became abruptly more lenient, and only about 5 percent became abruptly more severe. That is, most of the Readers who became more lenient or severe over the course of the Reading did so in a fairly systematic, progressive manner. The changes in rating behavior they exhibited were gradual, not sudden.
2. *Do different approaches to detecting differential severity/leniency produce similar results? How comparable are the indices of differential severity/leniency obtained from different approaches? Does each approach identify the same set of Readers as exhibiting differential severity/leniency?*
    - When we used the $Z_{SAI_{rc}}$ index to detect differential severity/leniency, the flag rates were much lower than when we used the SAI_{rc} index. The values of $F_{V(X)}$, however, were inconsistent, with these two DRIFT indices producing markedly different results.
3. *Are there any AP English Literature and Composition essay Readers who exhibit differential scale category use (i.e., change from using all the categories in the scoring guidelines to using fewer categories as the Reading progresses)? If so, at what point in the Reading is this Reader effect detectable? Are there some Readers who, throughout the entire Reading, do not use all the categories on the scoring guidelines?*
4. *Are there any AP English Literature and Composition essay Readers whose levels of accuracy change as the Reading progresses? If there are such Readers, at what point in the Reading do these changes become apparent?*
    - As a group, the Readers did not exhibit evidence of differential scale category use. There was weak evidence suggesting an overall increase in Reader accuracy over time. Specifically, the value of $Z_{SR-ROR_c, SR-ROR_b}$ increased from 0.13 to 0.67 from the second to the eighth time period. This corresponds to an increase in the single Reader–rest of the Readers correlation from about 0.82 to about 0.91 from Time 2 to Time 8.
    - Individually, about 5 percent of the Readers were flagged for differential scale category use (i.e., increased use of the central rating categories) at each time period. With the exception of a large jump from Time 1 to Time 3 (i.e., up to 10 percent), the opposite effect (i.e., decreased use of the central rating categories) was about half of that rate.
    - Individually, about 5 percent of the Readers were flagged for increased accuracy at each time period, but fewer than two Readers were flagged for decreased accuracy at each time period.

- Most of the Readers flagged for differential scale category use or increased accuracy were flagged only temporarily. That is, the effects were not sustained over several time periods. Only 5 percent of the Readers were flagged repeatedly (i.e., for at least three of the seven time-period comparisons) for differential scale category use, while only 2 percent of the Readers were flagged repeatedly for increased accuracy.

5. *Do different approaches to detecting differential scale category use in a Reader's ratings produce similar results? How comparable are indices obtained from different statistical approaches? Does each approach identify the same set of Readers as exhibiting differential scale category use?*
6. *Do different approaches to detecting differential accuracy/inaccuracy produce similar results? How comparable are the indices of differential accuracy/inaccuracy obtained from different approaches? Does each approach identify the same set of Readers as exhibiting differential accuracy/inaccuracy?*
   - The values of the single Reader–rest of the Readers correlation DRIFT index, $Z_{SR-ROR_c, SR-ROR_b}$, exhibited a moderately strong relationship with the values of the two Reader fit-based indices, $F_{fit-weighted}$ and $F_{fit-unweighted}$, and with the values of the Reader residuals–expected ratings index, $Z_{E-R_c, E-R_b}$—about −0.55 for the fit-based indices, and around 0.61 for the residuals-based index. The two fit-based indices were nearly perfectly consistent as indices of DRIFT. The $F_{V(X)}$ index did not correlate very highly with either the single Reader–rest of the Readers correlation or with the Reader fit-based DRIFT indices.
   - The Reader residuals–expected ratings DRIFT index, $Z_{E-R_c, E-R_b}$, exhibited a moderately strong relationship with $Z_{SR-ROR_c, SR-ROR_b}$ (r = 0.61) and with $F_{V(X)}$ (r = 0.43). However, its relationship with the Reader fit-based indices was weak (r was about −0.22).
   - Due to the small percentage of Readers flagged for DRIFT, the overall level of agreement between $Z_{SR-ROR_c, SR-ROR_b}$, $Z_{E-R_c, E-R_b}$, $F_{V(X)}$, and the values of the two fit-based indices, $F_{fit-weighted}$ and $F_{fit-unweighted}$, were fairly high. However, conditional flag rates were high only between the two fit-based indices. We observed moderately high conditional flag rates between $F_{V(X)}$ and $Z_{E-R_c, E-R_b}$, and low conditional flag rates between $F_{V(X)}$ and the two fit-based indices. All other conditional flag rates were close to zero.
7. *How do our depictions of DRIFT differ when we establish Reader connectivity via student responses to multiple-choice items versus via ratings of benchmark essays?*
8. *How do our depictions of DRIFT differ when we utilize Reader agreement indices versus when we utilize Reader accuracy indices?*
   - The methods for establishing connections between Readers that were based on agreement frameworks (i.e., the multiple-choice and the benchmark methods) produced DRIFT indices that were only weakly related to one another.
   - The accuracy and agreement methods for establishing connections between Readers produced mixed results in terms of DRIFT index comparability. Specifically, the DRIFT indices produced using the multiple-choice method correlated only weakly with the indices produced using the experienced Readers method. By contrast, the benchmark method and the experienced Readers method produced DRIFT indices that were highly correlated for detecting differential severity/leniency and moderately correlated for detecting differential scale category use and accuracy DRIFT.

Conclusion

Results from this study indicate that some Readers showed statistically significant changes in the levels of severity or leniency they exercised as the Reading progressed. The findings from the present study confirm those of other researchers (Bleistein and Maneckshana, 1995; Braun, 1988; Coffman and Kurfman, 1968; Hoskens and Wilson, 2001; Lumley and McNamara, 1995; Lunz and Stahl, 1990; Morgan, 1998; O'Neill and Lunz, 2000; Wilson and Case, 2000; Wood and Wilson, 1974) who found that, in some settings, Readers do not maintain a consistent level of severity over time. Unfortunately, we cannot determine whether the changes in severity/leniency that we observed in this AP Reading were due to Readers' incorrect initial expectations that they subsequently corrected, or vice versa (correct initial expectations that they erroneously changed).

In our study, a higher percentage of Readers became somewhat more lenient over time than became more severe. For any given time period, 12 to 21 percent of the Readers in the study rated significantly more leniently than they did during the first time period, while 2 to 11 percent of the Readers rated significantly more severely than they did during the first time period. In some of the more extreme examples, the average rating for some Readers increased by more than 1.5 rating scale points over the course of the Reading.

Fortunately, the majority of Readers who became more lenient (or severe) over time did so in a consistent

fashion. That is, once a Reader changed the level of severity exercised, that Reader continued to exercise the same level of severity through the remainder of the Reading. Few Readers drifted back and forth erratically, sometimes rating more leniently, while at other times rating more severely.

Additionally, some Readers in the study showed statistically significant changes in their levels of accuracy as the Reading progressed, while other Readers showed evidence of differential scale category use over time. However, not as many Readers showed these two types of DRIFT as showed DRIFT related to severity/leniency. For any given time period, 1 to 8 percent of the Readers rated more accurately than they did during the first time period, while 0 to 2 percent of the Readers rated less accurately than they did during the first time period. Also, for any time period, 3 to 7 percent of the Readers used fewer categories on the 9-point AP rating scale than they did during the first time period.

The encouraging news is that in the vast majority of cases, Readers exhibited no evidence of statistically significant changes over time in their levels of accuracy, or in their use of the scale categories. Most commonly, Readers who exhibited any evidence of these changes in their rating behavior did so for only one time-period comparison. This suggests that when Readers changed their rating behavior in terms of accuracy or scale category use, the changes were only temporary. That is, they did not continue that DRIFT behavior over time. However, it is important to point out that there were individual Readers who showed evidence of sustained DRIFT, which resulted in marked changes in their rating behavior that they then maintained across multiple time periods.

Which of the indices we studied are the most appropriate measures of DRIFT? Some Rasch indices seem to be clearer, less ambiguous indicators of DRIFT than other Rasch indices. For example, the ratios of the Reader mean-square fit indices ($F_{fit-unweighted}$ and $F_{fit-weighted}$) seem to be sensitive to both differential accuracy/inaccuracy and to differential scale category use. Hence, those two indices provide ambiguous information for diagnosing DRIFT and are therefore potentially less helpful than other indices for identifying Readers who are exhibiting a particular type of DRIFT.

The results from our study suggest that the following indices are most suitable for detecting Reader DRIFT:

- For the detection of DRIFT related to differential severity/leniency: Among the indices we studied, the $Z_{SAI_{rc}}$ statistic appears to be the most suitable measure of change in a Reader's level of severity or leniency over time.
- For the detection of DRIFT related to differential accuracy/inaccuracy: The single Reader–rest of the Readers correlation, Z_{SR-ROR_c}, $_{SR-ROR_b}$, provides the most sensitive measure of Reader accuracy.
- For the detection of DRIFT related to differential scale category use: The correlation between Rasch-based expected ratings and the residuals, $Z_{E-R_c,E-R_b}$, appears to be the most suitable index for detecting changes in Reader accuracy over time. There does not seem to be an adequate raw score index for detecting this form of DRIFT.[10] Factors other than changes in a Reader's use of the central rating scale categories influence the standard deviation of the raw ratings—the raw score index that researchers have most frequently employed in the past when they have tried to detect this type of DRIFT—rendering it an unsuitable index for this purpose. (Those confounding factors include the level of severity/leniency the Reader exercised, as well as how much variance in achievement the essays the Reader rated during a given time period showed. For example, did the set of essays the Reader evaluated in a particular time period contain an overabundance of essays of lower-achieving students, an overabundance of essays of higher-achieving students, or did the set contain a representative mix of essays from lower-, higher-, and average-achieving students? These factors can have a direct bearing on the standard deviation of the raw ratings and could lead one to mistakenly conclude that a Reader is inappropriately using the scale categories as a Reading progresses.)

A critical challenge for those who monitor Reader behavior is to find ways to identify in "real time" (i.e., while an AP Reading is in progress) those few Readers whose rating standards are drifting—those who are becoming more severe or lenient, those who are becoming more inaccurate, as well as those who are using fewer rating categories over time. Supervisory personnel can then take appropriate actions to attempt to get them back on track through additional training, recalibration, etc. In order to meet this need, the AP Program may want to consider carrying out real-time Many-Facet Rasch Measurement analyses of the rating data from its Readings. As this study has demonstrated, results from Many-Facet Rasch Measurement analyses could provide a valuable supplement to the information that the Reader Management System (RMS) currently provides during the Reading to aid the Table Leaders in identifying Readers who are experiencing DRIFT.

In order to determine whether any Readers are showing worrisome changes in their patterns of ratings as a Reading progresses, a data analyst must provide

[10] A possibility might be to model each Reader to have his or her own unique rating scale structure for each time point (Linacre, personal communication). It would then be possible to compare the rating scale structures for a given Reader across multiple time points.

trustworthy, time-sensitive measures of rating behavior. The analyst can get some sense of just how much more severe or lenient an individual Reader is at various points in time by comparing a given Reader's average ratings for two or more time periods (as is current Reader monitoring practice for the AP Readings). However, the analyst will not be able to determine how much the particular sample of essays that a Reader evaluated during a given time period influences that Reader's average rating. For example, if the average rating for Reader A is lower for Time 1 than it is for Time 2, there are two plausible explanations for this apparent change in rating behavior. Perhaps the set of essays that Reader A evaluated at Time 1 included more essays from lower-achieving students than the set of essays that Reader A evaluated at Time 2. If this were the case, the fact that Reader A tended to assign an overabundance of lower ratings at Time 1 would have been entirely appropriate, given that the majority of essays that this Reader evaluated during that time frame were essays from lower-achieving students. Chances are that other Readers evaluating this particular set of essays would also have given them low ratings. According to this explanation, Reader A has a lower average rating for Time 1 because the essays that Reader A evaluated were, by and large, from lower-achieving students. Following this line of reasoning, if Reader A had evaluated a set of essays that included more essays from higher-achieving students at Time 1, then Reader A would not have given an overabundance of lower ratings.

A second possible explanation for this apparent change in rating behavior is that Reader A tended to use the 9-point rating scale in a different manner at Time 1 than at Time 2. Perhaps the Reader's standards shifted. If this were the case, then when other Readers evaluated the set of essays that Reader A evaluated at Time 1, they would have assigned these essays higher ratings than Reader A gave them. When a Reader shows differences in the level of severity exercised across time periods, it is difficult to decide which of these explanations for the Reader's behavior is the correct one.

By using a Many-Facet Rasch Measurement approach to analyzing the rating data, a data analyst can gain needed insights to facilitate this determination. The output from a Many-Facet Rasch Measurement analysis contains a *fair average* rating for each reader. If the analyst were to run parallel Many-Facet Rasch Measurement analyses—one for each time period—the analyst could then compare an individual Reader's fair averages across time periods. The fair average is the average rating for each Reader once the MFRM has adjusted that average for the deviation of the essays in each Reader's sample from the overall essay average across all Readers for a given time period. The analyst could use the output from the analyses to pinpoint those Readers who showed evidence of becoming more severe or more lenient as the Reading progressed, even after taking into account the particular essays that that Reader evaluated during that time period.

The rating design employed in an AP Reading must meet certain design constraints if an analyst wants to be able to obtain interpretable results from a Many-Facet Rasch Measurement analysis. In the current AP rating design, Readers are nested within the three free-response questions, and there is no overlap among Readers. A single Reader rates each essay, and no Reader rates essays written for more than one free-response question. This incomplete block design (Ebel, 1951; Fleiss, 1981) characterizes a number of AP Readings, posing a challenge for all approaches to analyzing the data that are based on Item Response Theory, since the data matrices resulting from this type of design are typically sparse, containing much missing data.

If an analyst wants to draw comparisons among all Readers who are evaluating essays for a given free-response question, then the allocation of Readers to essays for that question must result in a network of links that is complete enough to connect all those Readers (Engelhard, 1997; Lunz, Wright, and Linacre, 1990).[11] When those in charge of a Reading carefully plan the rating design so that all Readers can be connected in a Many-Facet Rasch Measurement analysis, the analyst can then directly compare the Readers' severity measures that are produced as part of the output from the analysis.

If the connection among Readers in the rating design is not adequate, then there will be disconnected subsets of Readers in the analysis, which will be problematic for the measurement model as it carries out its estimation process. When there are disconnected subsets of Readers, Reader severity and level of student achievement are confounded, leading to ambiguity in the calibration of both the Readers and the students. That is, if the average rating for a Reader is lower than the average rating of other Readers (i.e. the Reader appears to be "severe" in comparison to others), the measurement model will not be able to determine whether the Reader tended to assign systematically lower ratings than other Readers, or whether the set of student essays the Reader evaluated tended to be lower in quality than other students' essays. In order for the measurement model to be able to disentangle differences in levels of Reader severity from differences in levels of student achievement, the various subsets of Readers need to be connected. If an analyst runs a Many-Facet Rasch Measurement analysis with disconnected subsets of Readers, then the analyst will only be able to compare the severity measures of Readers

[11] Disconnected subsets of Readers would make it difficult to compare levels of severity or leniency, but most fit statistics would probably be only minimally influenced by the disconnections.

that are included in the same subset—a much less useful outcome for quality control monitoring purposes than having output from an analysis that allows direct comparisons of all Readers included in the analysis.

In an earlier study (Engelhard and Myford, 2003), we demonstrated that it is possible to run a Many-Facet Rasch Measurement analysis on AP rating data by linking all the Readers through students' responses to the multiple-choice questions as a means of establishing the requisite connectivity in the rating design. In the present study, we experimented with a more direct method of establishing connectivity by introducing benchmark essays into the Reading (i.e., a set of essays that a group of highly experienced AP Readers previously rated in order to obtain a consensus rating for each essay). We say "more direct" because connecting Readers through students' responses to multiple-choice questions only allows the AP Program to obtain indirect measures of each Reader's performance (i.e., an indication of whether the ratings the Reader gives students' essays are in sync with the overall level of performance each student displayed on the multiple-choice section of the exam). On the other hand, the introduction of benchmark essays may raise questions about the generalizability of results from this study due to the fact that these essays were photocopied and appeared to be different from operational essays. Nevertheless, introducing benchmark essays into a Reading allowed us to monitor not only changes in each Reader's level of severity as the Reading progressed, but also changes in each Reader's level of accuracy over time. Hence, we were able to compare a Reader's performance to known standards of performance (i.e., the highly experienced Readers' consensus ratings of the benchmark essays).

In this study, we compared three different strategies for connecting Readers and found that, in some cases, the DRIFT indices we obtained were highly dependent upon the strategy we employed. When we implemented the multiple-choice strategy, the DRIFT indices we obtained showed only a weak relationship to the comparable DRIFT indices we obtained when we implemented the benchmark strategy. (For the multiple-choice strategy, we scaled the ratings of the essays, fixing student achievement measures at values derived from their responses to the multiple-choice questions. For the benchmark strategy, we scaled the ratings of the essays, allowing the benchmark essays rather than the students' responses to the multiple-choice items to serve as the links between Readers and between students.)

We suspect that the low correlations between these two sets of indices may reflect differences in how the MFRM handled the raw ratings. That is, when the measurement model established connections between Readers using the linking benchmark essays, it was able to disentangle the effects of Readers and students. By contrast, when the measurement model established connections between Readers using students' responses to the multiple-choice items, it did not have the information it needed from the commonly scored essays that would enable it to disentangle these effects. Consequently, the mean of the "unconnected" raw ratings derived from the implementation of the multiple-choice strategy differs by more than one rating scale point from the mean of the "connected" ratings (i.e., the expected ratings) derived from the implementation of the benchmark strategy.

Similarly, the DRIFT indices resulting from the implementation of the multiple-choice strategy showed only a weak relationship to the DRIFT indices resulting from the implementation of the experienced Readers strategy. (The experienced Readers strategy "fixes" the achievement measures of students whose essays are used as benchmarks.) Again, the lack of association between the two sets of indices may reflect differences in how the MFRM handled the raw ratings when we implemented these two strategies.

On a more positive note, some of the DRIFT indices produced using the benchmark and experienced Readers strategies showed more promise in terms of their potential comparability. Both of these strategies made use of the ratings that Readers gave to the benchmark essays to establish connections among Readers, which may account for the stronger associations between these two sets of DRIFT indices. Given these findings, the AP Program may want to consider seeding essays that highly experienced Readers have previously rated into the Reading at various points in time as a periodic check on whether each Reader is able to maintain an acceptable level of accuracy as the Reading progresses. Using this strategy would allow supervisory personnel to determine whether or not changes in Reader behavior identified in a Many-Facet Rasch Measurement analysis represented genuine cause for concern. Additionally, implementation of this strategy would provide a more direct basis for establishing connections among all the Readers.

References

Bleistein, C., & Maneckshana, B. (1995). *English literature and composition folder study* (Unpublished Statistical Report No. SR-95-39). Princeton, NJ: Educational Testing Service.

Braun, H. I. (1988). Understanding score reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics, 13*(1), 1–18.

Coffman, W. E., & Kurfman, D. (1968). A comparison of two methods of reading essay examinations. *American Educational Research Journal, 5*(1), 101–20.

College Board. (1999). *1999 Advanced Placement test analyses, forms 3VBP* (Unpublished Statistical Report). Princeton, NJ: Educational Testing Service.

Congdon, P. (1998). Unmodelled rater discrimination error. In M. Wilson & G. Engelhard Jr. (Eds.), *Objective measurement: Theory into practice* (Vol. 5). Stamford, CT: Ablex Publishing.

Draba, R. E. (1977). *The identification and interpretation of item bias* (Research Memorandum No. 26). Chicago: University of Chicago.

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika, 16,* 407–24.

Edgeworth, F. Y. (1890). The element of chance in competitive examinations. *Journal of the Royal Statistical Society, 53*, 460–75, 644–63.

Engelhard, G. Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5*(3), 171–91.

Engelhard, G. Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*(2), 93–112.

Engelhard, G. Jr. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33*(1), 56–70.

Engelhard, G. Jr. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement, 1*(1), 19–33.

Engelhard, G. Jr. (2002). Monitoring raters in performance assessments. In G. Tindal & T. Haladyna (Eds.), *Large-scale assessment programs for ALL students: Development, implementation, and analysis* (pp. 261–87). Mahwah, NJ: Lawrence Erlbaum Associates.

Engelhard, G. Jr., & Myford, C. M. (2003). *Monitoring faculty consultant performance in the Advanced Placement English Literature and Composition program with a many-faceted Rasch model* (College Board Research Report No. 2003-1). New York: The College Board.

Engelhard, G. Jr., Myford, C. M., & Cline, F. (2000). *Investigating assessor effects in National Board for Professional Teaching Standards assessments for Early Childhood/Generalist and Middle Childhood/Generalist certification* (ETS Research Report RR-00-13). Princeton, NJ: Educational Testing Service.

Fleiss, J. L. (1981). Balanced incomplete block designs for inter-rater reliability studies. *Applied Psychological Measurement, 5*, 105–12.

Garner, M., & Engelhard, G. Jr. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education, 12*(10), 29–51.

Good, P. I. (1999). *Resampling methods: A practical guide to data analysis*. Boston: Birkhauser.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Heller, J. I., Sheingold, K., & Myford, C. M. (1999). Reasoning about evidence in portfolios: Cognitive foundations for valid and reliable assessment. *Educational Assessment, 5*(1), 5–40.

Hoskens, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed response items: An example from the Golden State Examination. *Journal of Educational Measurement, 38*, 121–46.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (2003). Facets: Rasch Measurement Computer Program (Version 3.47) [Computer software]. Chicago: MESA Press.

Linacre, J. M., Engelhard, G. Jr., Tatum, D. S., & Myford, C. M. (1994). Measurement with judges: Many-faceted conjoint measurement. *International Journal of Educational Research, 21*(6), 569–77.

Lumley, T., & McNamara, T. F. (1995). Reader characteristics and reader bias: Implications for training. *Language Testing, 12*, 54–71.

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions, 13*, 425–44.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*(4), 331–45.

Manalo, J. R. (2002, April). *Detecting non-attending behaviors in questionnaire data using the Rasch rating scale model*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Morgan, R. (1998). *An examination of the impact of folder position and the reading day on the scoring of eight Advanced Placement Exams* (Unpublished Statistical Report SR-98-03). Princeton, NJ: Educational Testing Service.

Myford, C. M., Marr, D. B., & Linacre, J. M. (1996). *Reader calibration and its potential role in equating for the Test of Written English* (ETS Research Report No. 52). Princeton, NJ: Educational Testing Service.

Myford, C. M., & Mislevy, R. J. (1995). *Monitoring and improving a portfolio assessment system* (No. RR-94-05). Princeton, NJ: Educational Testing Service, Center for Performance Assessment.

Myford, C. M., & Wolfe, E. W. (2002). When raters disagree, then what: Examining a third-rating discrepancy resolution procedure and its utility for identifying unusual patterns of ratings. *Journal of Applied Measurement, 3*, 300–24.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*, 386–422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*, 189–227.

O'Neill, T. R., & Lunz, M. E. (2000). A method to study rater severity across several administrations. In M. Wilson & G. Engelhard Jr. (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 135–46). Stamford, CT: Ablex Publishing.

Paulukonis, S. T., Myford, C. M., & Heller, J. I. (2000). Formative evaluation of a performance assessment scoring system. In M. Wilson & G. Engelhard Jr. (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 15–40). Stamford, CT: Ablex Publishing.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495–502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*, 197–207.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413–28.

SAS Institute. (1999). SAS System for Windows (Release 8.00) [Computer software]: Cary, NC: Author.

Smith, R. M. (1996). Polytomous mean-square fit statistics. *Rasch Measurement Transactions, 10*, 516–17.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361–70.

Wilson, M., & Case, H. (2000). An examination of variation in rater severity over time: A study in reader drift. In M. Wilson & G. Engelhard Jr. (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 113–34). Stamford, CT: Ablex Publishing.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing, 4*, 83–106.

Wolfe, E. W. (1998, April). *Criterion-referenced rater monitoring through optimal appropriateness measurement*. Paper presented at the annual meeting of the American Educational Research Association, San Diego.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science, 46*(1), 35–51.

Wolfe, E. W. (2005). Identifying rater effects in performance ratings. In S. Reddy (Ed.), *Performance Appraisals: A Critical View* (pp. 91–103). Hyderabad, India: ICFAI University Press.

Wolfe, E. W., Chiu, C. W. T., & Myford, C. M. (2000). Detecting rater effects in simulated data with a multi-faceted Rasch rating scale model. In M. Wilson & G. Engelhard Jr. (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 147–64). Stamford, CT: Ablex Publishing.

Wolfe, E. W., & Gitomer, D. (2001). The influence of changes in assessment design on the psychometric quality of scores. *Applied Measurement in Education, 14*, 91–107.

Wolfe, E. W., Kao, C. W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication, 15*, 465–92.

Wolfe, E. W., & Moulder, B. C. (1999, April). *Examining differential reader functioning over time in rating data: An application of the multi-faceted Rasch rating scale model*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Wolfe, E. W., Moulder, B. C., & Myford, C. M. (2001). Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model. *Journal of Applied Measurement, 2*, 256–80.

Wood, R., & Wilson, D. (1974). Evidence for differential marking discrimination among examiners of English. *The Irish Journal of Education, 8*(1), 36–48

Wright, B. D. (1991). Diagnosing misfit. *Rasch Measurement Transactions, 5,* 156.

Wright, B. D. (1995). Diagnosing person misfit. *Rasch Measurement Transactions, 9*, 430–31.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.

Appendix

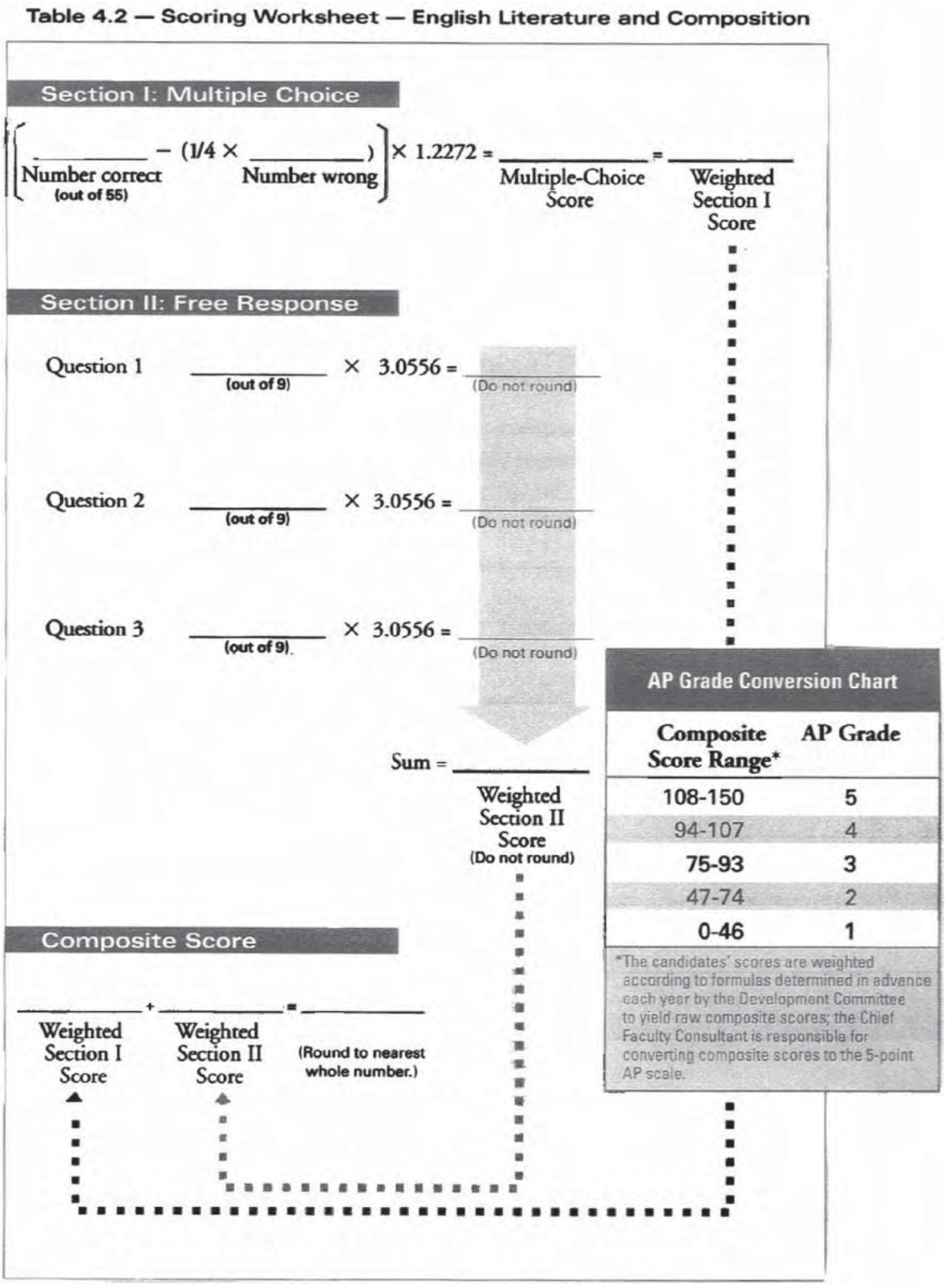


Figure A1. Description of scoring system (College Board, 1999, p. 72).

Exhibit A1

Scoring Guidelines for Question 1 from the 2002 AP English Literature and Composition Exam*

Question 1

Alain de Botton's *Kiss and Tell*

General Directions: This scoring guide will be useful for most of the essays that you read, but in problematic cases, please consult with your Table Leader. The score you assign should reflect your judgment of the quality of the essay as a whole. **Reward the writers for what they do well.** The score for an exceptionally well-written essay may be raised by one point above the otherwise appropriate score. In no case may a poorly written essay be scored higher than a three (3).

9–8:	These well-focused essays offer a persuasive interpretation of how Alain de Botton produces comic effect in his dramatic depiction of a scene in which Isabel unexpectedly discovers that her parents are in the same theatre as she and her new boyfriend. Specifically, the writers of these essays identify techniques and analyze how the author uses them to create comic effect. These essays make apt and specific references to the passage, effectively analyzing the nature of the comic effect that the author derives from the situation itself, from the thoughts of Isabel and her conversation with her parents, and from the relationship between daughter and parents. Though these essays may not be error-free, they are perceptive in their analysis of the comic effect and demonstrate writing that is clear and precise. Generally, the nine (9) essays reveal a more sophisticated analysis and a more effective control of language than do the essays scored an eight (8).
7–6:	These competent essays offer a reasonable interpretation of how Alain de Botton produces a comic effect. The writers identify the techniques and analyze how the author employs them. Although not as convincing or as thoroughly developed as those in the highest range, these essays demonstrate the writer's ability to express ideas with clarity, insight, and control. Generally, the seven (7) essays present a more developed analysis and a more consistent command of the elements of effective composition than do essays scored a six (6).
5:	These essays offer a plausible interpretation of how Alain de Botton achieves comic effect, but they often respond to the assigned task with a simplistic Reading of the passage. They often rely on paraphrase, but the paraphrase will exhibit some analysis, implicit or explicit. The discussion of the techniques may be slight and/or formulaic. These writers demonstrate some control of ideas, but the writing may be flawed by surface errors that do not create confusion for the reader.
4–3:	These lower-half essays offer a less than thorough treatment of the task. The analysis of the techniques used for comic effect may be partial, unconvincing, or irrelevant. These essays may rely on mere summary or be marked by observation rather than by analysis. The writing often demonstrates a lack of control over the conventions of composition: inadequate development of ideas, an accumulation of errors, or a focus that is unclear, inconsistent, or repetitive. Essays scored a three (3) may contain significant misreadings and/or distracting errors in grammar and mechanics.
2–1:	These essays compound the weaknesses of the papers in the 4–3 range. They may demonstrate an inability to explain how a comic effect is achieved, or even fail to recognize the comic effect. They may also be unacceptably brief or incoherent. The writing may contain pervasive errors, which interfere with understanding. Although some attempt has been made to respond to the question, the writer's assertions are presented with little clarity, organization, or support. Essays scored a one (1) contain little coherent discussion of the passage.
0	Indicates a response with no more than a reference to the task.
—	Indicates a blank paper or completely off-topic response.

* The scoring guidelines for Question 1 are taken from the *2002 AP English Literature and Composition Released Exam.*


CollegeBoard
connect to college success™